DOCUMENT RESUME

ED 072 421 CS 000 367

AUTHOR	Gadway, Charles J., Ed.
TITLE	Reading and Literature: General Information Yearbook.
INSTITUTION	Education Commission of the States, Denver, Colo. National Assessment of Educational Progress.
SPONS AGENCY	National Center for Educational Statistics (DHEW/OE), Washington, D.C.
REPORT NO	PR-02-GIY
PUB DATE	May 72
NOTE	96p.
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	Data Analysis; *Evaluation Methods; *Evaluation Techniques; *Literature; Measurement; *Measurement Goals; *Reading; Testing
IDENTIFIERS	NAEP; *National Assessment of Educational Progress

ABSTRACT

This report is the first volume of a projected series of "General Information Yearbooks" to be updated and published annually by the National Assessment of Educational Progress and is designed to aid the readers of the National Assessment's reports to understand the philosophies and goals of the project, its methods for developing objectives and exercises, and its techniques for collecting, processing, and describing data. This yearbook is oriented toward the reading and literature subject areas and the procedures and techniques relevant to the assessment year 1970-71. (Author/TO)

NATIONAL ASSESSMENT OF EDUCATIONAL PROGRESS

A Project of the Education Commission of the States

Robert W. Scott, Governor of North Carolina, Chairman, Education Commission of the States

Wendell H. Pierce, Executive Director, Education Commission of the States

James A. Hazlett, Administrative Director, National Assessment

Assessment Reports

#1	Science: National Results	July, 1970
#2a	Citizenship: National Results -- Partial	July, 1970
#2	Citizenship: National Results	November, 1970
#3	Writing: National Results	November, 1970
#4	Science: Group Results A	April, 1971
#5	Writing: Group Results A	April, 1971
#6	Citizenship: Group Results A	July, 1971
#7	Science: Group Results B	December, 1971
#8	Writing: National Results -- Writing Mechanics	February, 1972
02-R-20	Reading: Selected Exercises	May, 1972
02-GIY	Reading and Literature: General Information Yearbook	May, 1972

The project reported herein was performed pursuant to a grant from the National Center for Educational Statistics of the U.S. Office of Education, Department of Health, Education, and Welfare. However, the opinions expressed herein do not necessarily reflect the position or policy of the U. S. Office of Education, and no official endorsement by the U. S. Office of Education should be inferred.

This report was made possible in part by funds granted by Carnegie Corporation of New York and the Ford Foundation's Fund for the Advancement of Education. The statements made and views expressed are solely the responsibility of National Assessment of Educational Progress, a project of the Education Commission of the States.

Education Commission of the States
Suite 300, 1860 Lincoln Street
Denver, Colorado 80203

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY

ED 072421

NATIONAL ASSESSMENT OF EDUCATIONAL PROGRESS

A Project of the Education Commission of the States

REPORT 02-GIY

READING AND LITERATURE: GENERAL INFORMATION YEARBOOK

May, 1972

CS 000367

NATIONAL ASSESSMENT OF EDUCATIONAL PROGRESS

James A. Hazlett
Administrative Director

J. Stanley Ahmann
Staff Director

George H. Johnson
Associate Staff Director

This report is the product of many persons working together. Editor of this volume is Charles J. Gadway, Assistant to the Director, Department of Research and Analysis, National Assessment of Educational Progress.

Contributions were also made by Professor Frederick Mosteller of Harvard University, James Chromy of Research Triangle Institute, R. Paul Moore of Research Triangle Institute, and John O'Neill of Measurement Research Center.

Staff support was provided by:

Exercise Development Department
Information Services Department
Operations Department
Research and Analysis Department
Utilization/Applications Department
Data Processing Services Department (ECS)

TABLE OF CONTENTS

Foreword	. . . . .	i
Chapter 1:	General Introduction	1
Chapter 2:	Developing Objectives and Exercises . . .	4
Chapter 3:	Definition of National Assessment Groups .	12
Chapter 4:	National Assessment Samples	20
Chapter 5:	Administering the Assessment	48
Chapter 6:	Scoring of Exercises	55
Chapter 7:	Processing the Data	58
Chapter 8:	Describing National Assessment Data . . .	60
Chapter 9:	Overview of Selecting Exercises to be Reported	63
Chapter 10:	Details of Selecting Exercises to be Reported	65
Chapter 11:	Overview of Inferring Population Facts From Sample Data	73
Chapter 12:	Details of Inferring Population Facts From Sample Data	75
Chapter 13:	Themes	81
Chapter 14:	The Merits and Weaknesses of Adjustments (Including Balancing)	83

LIST OF EXHIBITS

Exhibit	Page
Exhibit F-1	v
Exhibit 4-1	26
Exhibit 4-2	28
Exhibit 4-3	33
Exhibit 4-4	33
Exhibit 4-5	34
Exhibit 4-6	35
Exhibit 4-7	36
Exhibit 4-8	41
Exhibit 4-9	41
Exhibit 4-10	42
Exhibit 4-11	43
Exhibit 4-12	44
Exhibit 4-13	47
Exhibit 9-1	68
Exhibit 9-2	70

List of Exhibits (Continued)

Exhibit	Page
Exhibit 9-3	71
Exhibit 9-4	72

FOREWORD

This is the first volume of a projected series of General Information Yearbooks to be updated and published annually by the National Assessment of Educational Progress. These yearbooks are designed to aid the readers of National Assessment's reports to understand the philosphies and goals of the National Assessment project, its methods for developing objectives and exercises, and its techniques for collecting, processing, and describing data. While these yearbooks will contain much general philosophical information about National Assessment, each volume will be oriented toward the subject areas, procedures, and techniques relevant to the assessment year currently being reported. This volume, therefore, is oriented to the year 02 assessment of Reading and Literature (assessed 1970-71). The General Information Yearbook was designed to complement the modified reporting format beginning with the year 02 reports.

The Reporting Format

The results for the year 01 assessment (Science, Citizenship, and Writing) have been reported in what might be called a "phase" format. Each phase volume treated a different aspect of the results for all the exercises within a given subject area. For example, within the subject area Science, the phase I report gave the national results for all the Science exercises; the phase IIA report presented results for the regions of the country, the sexes, and sizes of community; and the phase IIB report presented results for races, levels of parental education, and sizes-and-types of community.[1] A major disadvantage of this format is that it is difficult for our readers to follow the results through the various phases.

Because we feel that National Assessment results should be easily understood and both meaningful and useful to educators and other concerned persons, we have adopted an alternative reporting format with the following qualities: (1) the data are presented in a larger number of relatively small volumes so that the reader will not be overwhelmed by the sheer size of a report; (2) each volume contains exercises which cluster

---

[1]These groups are defined in chapter 3.

together in a way that is meaningful to educators and scholars in the relevant subject area; and (3) along with each exercise the reports give all data[2]--national and group--relevant to that exercise.

We believe that the clusters of exercises most meaningful to educators and scholars are what we call themes. A theme is a set of exercises which share a common idea but which may require diverse behavioral responses from respondent.[3] For each subject area in the year 02 assessment, there is a summary volume which provides all information specifically relevant to the subject area (including the themes and objectives for the area) and the general trends that appear in the data. A separate volume is devoted to each theme in which the data for the specific exercises within the theme are given along with a summary of the data for the theme.

Released and Unreleased Exercises

Not all the administered exercises are selected to be released for publication. The percentage of exercises to be released varies from subject area to subject area. The unreleased exercises are being withheld to be readministered in the future in order that the results of the two administrations can be compared. Only the exercises selected for immediate publication are discussed in the text of the theme volumes. Chapters 9 and 10 discuss the selection of exercises to be reported.

The released exercises have been assigned a five-digit identification number. The first digit identifies the theme which the exercise represents. The second and third digits identify the number of the exercise within the theme, and the fourth and fifth numbers identify subparts within exercises where applicable.

---

[2]Except for certain data to be given in special research volumes.

[3]Chapter 13 gives the rationale of themes, the development of themes within a subject area, and an expanded definition of what themes are and their function.

Volume Numbering System

Beginning with the year 02 reports, a new volume numbering system is being inaugurated which reveals the assessment year, subject area, and volume content. A volume number, therefore, consists of three parts. The first part is a two-digit number referring to the assessment year, for example, 02 for the second assessment year. The second part is an initial letter or letters for the subject area as follows:

A	= Art
COD	= Career and Occupational Development
C	= Citizenship
L	= Literature
MA	= Mathematics
MU	= Music
R	= Reading
S	= Science
SS	= Social Studies
W	= Writing

The third part is a two-digit number indicating both the content of the volume (summary, theme, or research) and the order of publication as follows:

00	= summary for a subject area
01	= observed results for theme 1
.	through
19	= observed results for theme 19
20	= numbers for research volumes such as
.	reports of balanced results and
29	others

For example, the Reading Summary is 02-R-00, Reading theme 5 is 02-R-05, and Reading "balanced results" is 02-R-20. Music theme 2 is 03-MU-02.

Limitations of the Data

The National Assessment of Educational Progress was created to provide educators, scholars, and lay persons concerned with education with data regarding the educational

acnievements of various groups[4] of young Americans in 10 subject areas.[5] Within the limits of error due to measurement and sampling error,[6] the obtained data as presented in National Assessment reports accurately describe the educational achievements of these groups as they actually exist in the real world. These obtained data portray the real problems facing education--improving the educational achievements of various groups of students.

When the data show that a group has achieved either above or below the nation as a whole, one must exercise extreme caution in attributing causation to these obtained differences. National Assessment is not intended to provide reasons for differences; its purpose is to describe such differences if they exist. Many factors may affect an individual's ability to give acceptable responses to exercises in the assessed subject areas. Consider for example a hypothetical group whose achievement is well above the national average. Most members of the group may attend schools which have excellent physical facilities and high quality faculties, belong to high socio-economic families, have parents with a high level of education, come from homes with many reading materials and so on. All these factors could contribute to the group's high level of achievement while membership in the group itself may contribute very little or nothing.

The name of a group is merely a categorical lable. When we look at the data for a given group, therefore, we cannot say any difference in achievement between that group and the nation as a whole is attributable solely to membership in that group. In other words, a group must not be construed as necessarily being the cause or even being a cause for the differential achievement between that group and the nation as a whole.

Often members of groups like the hypothetical group described above are distributed in such a way that a disproportionately large percentage of them are also members of other groups which contribute to the group's high (or low) level of

---

[4]These groups are defined in chapter 3.

[5]See footnote 1, chapter 1.

[6]See chapter 11.

achievement. The data obtained from these groups do not allow one to evaluate the effectiveness of the educational process on these groups apart from the advantageous (or debilitating) factors. A statistical procedure called balancing adjusts for the disproportionate distribution of group members in other categories or groups for which there are adequate data available. This procedure gives the achievement data for the group in question that would have been obtained had the members of the group been distributed proportionately across these other categories or groups. National Assessment data, balanced for disproportionate representation, are presented in a special research volume. Again extreme caution must be exercised in interpreting the balanced data. The balanced data still reflect many extraneous factors not assessed by National Assessment and, therefore, are still not "pure" measures of the achievements of the group in question. Even with balanced data, a group must not be construed as necessarily being the cause or even as being a cause for the differential achievement between that group and the nation as a whole.

* * * * * * * * * * * * * * *

Exhibit F-1 gives the number of respondents to whom each exercise package was administered at each age level. A horizontal line separates the group-administered packages (above the line) from the individual-administered packages. Packages are unique at each age level, therefore, package 5 (for example) is not the same at age levels 9, 13, 17, and adult.

Exhibit F-1

Pkg. No.	9	13	17	Adults
1	2610	2602	2405	1241
2	2518	2543	2414	1259
3	2595	2550	2361	1251
4	2589	2559	2428	1254
5	2625	2537	2332	1242
6	2595	2550	2308	1226
7	2558	2509	2394	-
8	2548	2542	2354	-
9	2573	2540	2343	-
10	2201	2583	2322	-
11	2210	2534	2125	-
12	2195	2591	2145	-
13	-	2547	-	-
14	-	2197	-	-
15	-	2199	-	-

CHAPTER 1

GENERAL INTRODUCTION

History and Purpose of National Assessment

By the early 1960's many billions of dollars were being invested annually in the formal education of our young people. The only available measures of educational quality resulting from this investment had been based upon inputs into the educational system such as teacher-student ratios, number of classrooms, and number of dollars spent per student. The tenuous assumption had been that the quality of educational outcomes--what students actually learn--was directly related to the quality of the inputs into the educational system. No significant direct assessment of educational outcomes had been made. The typical state-administered or school-administered achievement tests, which provided scores whereby one student could be compared with others, were useful for categorizing students; but they provided very little information about what students were actually learning.

This insufficiency of information became the concern of Francis Keppel, United States Commissioner of Education (1962-1965), who initiated a series of conferences to find ways in which it might be overcome. In 1964, a result of these conferences, John W. Gardner, president of the Carnegie Corporation, asked a distinguished group of educators and lay persons to form the Exploratory Committee on Assessing the Progress of Education (ECAPE). This committee, chaired by Dr. Ralph W. Tyler, was to examine the possibility of conducting an assessment of educational attainments on a national basis.

After much study, ECAPE deemed that it was feasible to inaugurate an assessment project to fill the information gap regarding the quality of educational outcomes by periodically assessing the knowledges, understandings, skills, and attitudes in 10 subject areas[1] at four age levels (9,13,17, and adult--ages 26-35). The project began its first assessment of the subject areas Science, Citizenship, and Writing in the Spring

---

[1] Art, Career and Occupational Development, Citizenship, Literature, Mathematics, Music, Reading, Science, Social Studies, and Writing.

of 1969. Later that same year, the project came under the auspices of the Education Commission of the States and was named the National Assessment of Educational Progress (NAEP).

For the first time, there would be a direct measure of educational outcomes which could be utilized by school systems to improve the educational process. Since NAEP is to be an ongoing project, it will eventually be able to assess changes in these knowledges, understandings, skills, and attitudes to determine any changes in educational outcomes.[2]

The Philosophy of Assessment

The typical achievement test measures people. Individuals respond to a number of questions or tasks (items), and a score is determined for each individual. On the basis of many such individual scores, one individual can be compared with any other or with the mean (average) score for the entire group. Little attention is paid to the specific knowledges or skills possessed either by the individual or by the entire group.

National Assessment measures knowledges, understandings, skills, and attitudes. We do not obtain scores on individuals. Rather, we obtain the percentages of individuals (at the four age levels mentioned above) in the nation as a whole and certain groups[3] who are able to respond acceptably[4] to exercises which reflect specific knowledges, understandings, skills, or attitudes.

National Assessment respondents respond to a set of questions or tasks much the same as they would on a typical achievement test. Since we do not obtain individual scores for

---

[2]This section was adopted from What is National Assessment by Dr. Frank Womer and The National Assessment Approach to Exercise Development by Drs. Carmen J. Finley and Frances S. Berdie and may be obtained from: National Assessment of Educational Progress, Public Information Department, 300 Lincoln Tower, 1860 Lincoln St., Denver, Colorado 80203.

[3]National Assessment's groups are listed and defined in chapter 3

[4]Chapter 7 discusses the ways in which National Assessment data are described.

these sets of questions or tasks, we call the sets "packages" rather than tests. We also call the questions or tasks "exercises" since they allow the respondent to demonstrate whetner or not he possesses the knowledge, understanding, skill, or attitude to respond acceptably to the exercise.

All standardized achievement test items ideally are of medium difficulty to assure maximal discrimination. Assessment exercises are equally represented by easy, medium, and hard difficulty levels, i.e., we expect certain knowledges, understandings, skills and attitudes to be possessed by a larger percentage of individuals than others.

CHAPTER 2

DEVELOPING OBJECTIVES AND EXERCISES

Criteria for Developing Objectives and Exercises

One of the most important features of National Assessment's developmental work is the formulation of educational objectives which govern the direction of the assessment in any given subject area. These objectives define a set of goals which are agreed upon by both laymen and educators as desirable directions in the education of young people. For National Assessment, these educational objectives must be acceptable to three major groups of people. First, the objectives must be considered important by scholars in the discipline of a given subject area. Second, objectives should be acceptable to most educators and be considered important teaching goals in most schools. Finally, and most uniquely, lay citizens interested in education must agree that the objectives are important for young people to attain and that these objectives are of value in modern life.

People from each of these three groups, representing different geographical sections of the country as well as different viewpoints, are brought together to help formulate and review National Assessment objectives. This does not mean that National Assessment objectives are the only ones with which all educators and lay people agree, but rather that the final set of objectives for any area is a summarization of the feelings of a cross section of the scholars, educators, students and lay citizens in this country. In addition, the final set of objectives for any area includes objectives which express minority viewpoints as well as those objectives which a majority of the people in these three groups consider important. To keep the objectives current and have them meet increasing demands for information pertinent to the evaluation of American education, these objectives are reviewed and revised when a new assessment cycle begins.

Once these educational objectives are identified, exercises are developed that will be used to measure the extent to which young people are achieving them. The term "exercise" is deliberately used to distinguish National Assessment materials from standardized tests. Most standardized tests are normative; that is, they seek to evaluate the individual with respect to some group of people. The goal of National Assessment, however, is to measure the knowledges, understandings, skills, and attitudes of large groups of people and determine their level of attainment. The results of assessment are reported in terms

of individual exercises; for example, "90% of all 17-year-olds know the name of their states, but only 20% know the name of their Representative in Congress." In addition to providing information on the achievement of 9-, 13-, 17-year-olds and Adults, ages 26 to 35, for the nation as a whole, National Assessment also provides results by regions of the country, size and type of community, sex, parental education, and color.[1]

Because of the uniqueness of the National Assessment approach to information gathering, the following criteria have been identified as guidelines for the development of National Assessment exercises.

Content Validity. Every exercise must be a direct measure of some important knowledge, understanding, skill, or attitude that reflects one or more objectives in a subject area. An exercise must be meaningful, and directly related to the objective that it is intended to fulfill. In practice, then, an exercise has content validity if it makes sense to a reader who sees it together with an objective and says, "Yes, this is a good measure of the knowledge or skill called for by the objective." The chief criterion for determining content validity is that an exercise pass extensive reviews by subject matter and lay groups (described later under procedures for exercise development). But to further insure that its exercises are meaningfuland relevant. National Assessment involves young people at the ages being assessed in both the formation and review of its exercises.

Clarity. Every exercise must be easy to understand so that the respondent knows what he is being asked to do. This means that the directions and formats which go with the exercises must be simple enough for anyone to understand and the vocabulary, phraseology, and length of sentences must not be confusing.

Type of Exercises and Formats. National Assessment instructs developers of its exercises to use that format which provides the best and most direct measure of the objective being assessed, and to create exercises which obtain actual samples of a young person's skills. In doing this, they are also encouraged to use, whenever possible, individual interviews, group discussions, or observations of group tasks to supplement the usual paper and pencil exercises. NAEP asks its

---

[1]Groups within these categories are defined in chapter 6.

exercise writers to experiment with unconventional stimulus materials such as pictures, tapes, films or practical everyday items in order to heighten interest and bring greater variety to the format. Most importantly, NAEP emphasizes the necessity for developing a pool of exercises which will reflect the cultural pluralism that exists in America and which will have meaning for all groups in our society.

Difficulty Level. Since National Assessment does not intend to rank people in order, it is not appropriate that its exercises be of medium difficulty as usually they are in standardized tests. The intent of National Assessment is instead to be able to describe the knowledges, understandings, skills, and attitudes of the most and least able young people as well as those of the average young person. Therefore, it develops easy,medium and difficult exercises which are aimed at each of these three groups.

Overlap Between Ages. In National Assessment an exercise which "overlaps" is one that is appropriate for more than one age level. An exercise generally overlaps between two adjacent age levels, but may overlap three or even all four age levels. The development of these overlapping exercises has been encouraged because of the interesting comparisons that can be made by giving the same exercise to two or more age levels.

Instructions for Scoring. Particularly those for open-ended types of exercises, determine what information is finally reported. Because of this, National Assessment requires its exercise writers to submit a detailed explanation of how to score and report each exercise. When the exercises are administered on a tryout basis to an actual sample of respondents, the results are used to check the accuracy and completeness of these scoring and reporting instructions.[2]

Procedures for Developing Objectives and Exercises

During the first years of the project most of the development of objectives and exercises was handled exclusively by large contracting organizations such as Educational Testing Service, Science Research Associates, and American Institutes for Research. The National Assessment staff during this time was small, and National Assessment's role in exercise development was to act as a guide rather than a direct supervisor. However, since 1970, there has been a considerable increase in

---

[2]Chapter 6 gives a detailed account of the scoring procedures.

the size of the National Assessment staff, and newly created interdepartmental teams have brought about a greater coordination of developmental activities. In some areas, subject matter specialists have assisted the Assessment in the preparation of objectives and exercises. The developmental process currently in use incorporates previous developmental experiences, and provides for closer monitoring by National Assessment staff of its contractors (individual specialists as well as large testing organizations). This development process is divided into five major phases:

Phase A - Development (or revision) and Review of Objectives and Prototype Exercises

Phase B - Preparation of Exercises and Exercise Development Tryouts

Phase C - Review and Revisions of Exercises

Phase D - Field Testing and Review of Tryout Results

Phase E - Final Reviews and Selection

Following is a description of each of these five phases.

Phase A: Development (or revision) and Review of Objectives and Prototype Exercises

This phase allows for reconsideration and possible revision of objectives in a previously assessed area for the development of objectives in a new area. A variety of inputs are sought in this early phase. The developer of material reviews the literature in the field and collects appropriate information from universities, state departments of education, school districts, and objectives exchanges around the country in order to determine the latest ideas in objectives development. Subject matter and objectives specialists formulate new objectives or revise the existing ones. These specialists, representing a broad cross section from around the country, are selected from elementary and secondary schools (public, private, and parochial) as well as from colleges, universities, and other professional organizations. All of these various inputs lead to the drafting of a set of objectives.

The next step is to hold a review by a separate group of subject matter specialists to critique the new or reformulated objectives. Again, as in the case of the specialists who helped formulate the objectives, there is broad professional representation. The specialists evaluate the objectives for

content validity, appropriateness, relevance, and reportability. If three reviewers think it necessary, the objectives may be further revised and refined.

In the process described above, both subject matter specialists and educators participate. The next step is to seek the reactions of lay citizens. Lay people from around the country are invited to attend a review of the objectives at which numerous panels are convened with each panel having representation from different occupations, socio-economic backgrounds, and areas of the country. Comments from these panels are used as the basis for further refinement of the objectives.

While the objectives approach their final form, exercise writers develop prototype exercises. A prototype exercise is one which serves as a model for the development of other exercises; it must be a measure of an objective, it must be clearly stated and include a rationale, directions to administrators, scoring instructions, a key or sample of acceptable and unacceptable responses, and a scheme for reporting of results. The ideas for prototype exercises originate with subject matter specialists representing different specialties within the field and people experienced at different levels of performance. When they are collected, they are placed in an exercise format and given a small scale tryout. The results from the tryouts and the prototype exercises themselves are then reviewed by an Exercise Development Advisory Group consisting of leading education and measurement specialists. This group either recommends that the prototypes be further refined or that the production of exercises should begin, with the prototype exercises guiding the contractor in the types of exercises that should be produced.

Phase B: Preparation of Exercises and Exercise Development Tryouts

The number of exercises produced depends upon the number of total minutes that can be administered in the assessment for a particular area and the number of exercise minutes that already exist (unreleased exercises from a previous assessment or exercises in a previous pool). The development contractor uses as varied a range of writers as feasible, including some of the writers who helped develop the prototype exercises. Once he has a pool of new exercises, the contractor conducts a limited tryout to obtain some samples of actual responses. Ideally, this tryout should be conducted by the exercise writers so that they can obtain some data on the clarity of the task, scoring categories, and administrative feasibility. Another purpose of the tryout is to provide information for lay and subject matter reviewers about the results the exercises will yield and the

possible problems that will have to be investigated before the actual assessment.

Phase C: Review and Revisions of Exercises

National Assessment submits the exercises to both subject matter specialists and lay people for their review and criticism. A panel of such reviewers is formed with the following considerations in mind:

There should be

1. Representation from all parts of the country.
2. Representation from different specialties within the discipline (academicians and other specialists working in professional capacities in the field)
3. Representation from organizations or projects concerned with the area
4. Public, private and parochial school people (classroom teachers, curriculum specialists, administrators) - both elementary and secondary levels.
5. Persons knowledgeable about minority groups
6. Persons knowledgeable about low socio-economic problems
7. Students
8. A balance between male and female participants

For the selection of lay reviewers, the following criteria apply:

1. There should be geographical representation and people from various types of communities, i.e., towns, rural areas, big cities, etc.
2. Representation from different national organizations
3. Representation from different occupations
4. Persons knowledgeable about minority groups

5. Persons knowledgeable about low socio-economic problems

6. A balance between male and female participants

The subject matter specialists in their review of the exercises are primarily concerned with content validity, relevance, administrative feasibility, completeness of scoring and reporting instructions; the lay reviewers are primarily concerned with appropriateness and relevance.

Phase D: Field Testing and Revision of Exercises

After the exercise pool has been revised according to suggestions made by the reviewers there is a full scale tryout involving representatives of the community, socio-economic, racial, sex, and regional groups that are found in the actual assessment. The purpose of the tryout is to give some indication of how each exercise will probably function in an assessment, and to provide information that can be used to improve administrative and scoring instructions.

Following the tryout, subject matter specialists review the results to make sure that the scoring guides reflect actual responses and will provide important and desirable information.

Phase E: Final Reviews and Selection

At this point all the exercises have been through thorough reviews by both subject matter specialists and lay citizens, and are candidates for the assessment. However, the number of exercises developed always exceeds the number that can actually be used in an assessment. Specialists in the area rate each exercise according to its importance and quality, and these ratings are then reviewed by a committee consisting of subject matter and measurement specialists plus National Assessment and contractor staff members who have been involved in the development of the area. This committee designates the exercises to be administered in the actual assessment on the basis of the quality ratings and reporting needs (coverage of objectives). The selected exercises are then forwarded to the United States Office of Education to be checked for any infringement of privacy on the part of respondents or possible offensiveness.

Additional Development Work

During preparation for the packaging of exercises and printing of assessment booklets (packages) in any of the areas, exercises are closely examined by members of both the National

Assessment Team as well as contractors responsible for the packaging, printing, and field administration functions. In some cases, administrative instructions are added to facilitate administration in the field.

Development of Objectives and Exercises Year 02 - Reading and Literature

The development of objectives for Reading was the contract responsibility of Science Research Associates, and for Literature this responsibility was handled by Educational Testing Service. In both areas this was the first development of objectives, and this development followed the process described under Phase A of the previous section on Procedures for Developing Objectives and Exercises.

The development of exercises in Reading was initially handled by Science Research Associates. However, following the subject matter review of the Reading exercises in Phase C it was decided that the existing pool of exercises needed to be supplemented with additional exercises. Independent writers were contracted to develop these additional exercises which were in turn reviewed by subject matter specialists and lay citizens and then added to the existing pool of exercises. In Literature the development of exercises was completed by Educational Testing Service. Except for Phase B exercise development tryouts, procedures for the development of both Reading and Literature exercises followed the process described in the section entitled: Procedures for Developing Objectives and Exercises. The Phase B exercise development tryouts were added as a stage in the developmental process following year 03 work to permit the detection of deficiencies in exercises early in the developmental process.

CHAPTER 3

DEFINITION OF NATIONAL ASSESSMENT GROUPS

We report results for four age levels which represent the completion or near completion of significant educational levels by most individuals at each respective age:[1] 9-year-olds, most of whom have completed the primary grades (kindergarten through grade 3); 13-year-olds, most of whom are nearing the completion of the intermediate grades (grade 4 through grade 8); 17-year-olds, most of whom are nearing the completion of the secondary grades (grade 9 through grade 12); and Adults aged 26-35, most of whom have completed all formal education.[2] Within each of these age levels, we report results[3] for the nation as a whole and for geographic regions of the country, colors, sizes-and-types of community, and levels of parental education as follows:

Characteristic	Group
Sex	Male Female
Region of the Country	Northeast Southeast Central West
Color	Black White Other

---

[1]Some individuals at each age level have not achieved the same educational status as the majority.

[2]The actual grades included in the primary, intermediate, and secondary classifications may vary among school systems.

[3]Chapter 7 explains how National Assessment results are described.

Characteristic	Group
Size-and-Type of community	Extreme Rural Extreme Inner City Extreme Affluent Suburb Inner City Fringe Suburban Fringe Medium City Small City
Level of parental education	No High School Some High School Graduated High School Post High School

For the year 02 assessment, these age levels and groups are defined as follows:

Age Level

9-year-olds: all individuals born from January 1, 1961, through December 31, 1961.

13-year-olds: all individuals born from January 1, 1957, through December 31, 1957.

17-year-olds "in school": all individuals enrolled in secondary school and born from October 1, 1953, through September 30, 1954.

17-year-olds "out-of-school": all individuals not enrolled in secondary school during March, 1970, and born from October 1, 1952, through September 30, 1953; and all individuals not enrolled in secondary school during January, 1971, and born from October 1, 1953, through September 30, 1954.

Adults: all individuals born from April 1, 1935, through March 31, 1945.

Region of the Country

National Assessment uses the same regional division that is used by the Office of Business Economics, Department of Commerce.

Northeast	Southeast	Central	West
Connecticut	Alabama	Illinois	Alaska
Delaware	Arkansas	Indiana	Arizona
District of Columbia	Florida	Iowa	California
Maine	Georgia	Kansas	Colorado
Maryland	Kentucky	Michigan	Hawaii
Massachusetts	Louisiana	Minnesota	Idaho
New Hampshire	Mississippi	Missouri	Montana
New Jersey	North Carolina	Nebraska	Nevada
New York	South Carolina	North Dakota	New Mexico
Pennsylvania	Tennessee	Ohio	Oklahoma
Rhode Island	Virginia	South Dakota	Oregon
Vermont	West Virginia	Wisconsin	Texas
			Utah
			Washington
			Wyoming

Color

Individuals were classified as Black, White, or Other on the basis of information they provided. Results are given for Blacks and Whites only. The number of individuals classified as Other was too small to produce reliable results.

Size-and-Type of Community

The size-and-type of community category (STOC) integrates three extreme types of community (TOC), each composed of approximately 10% of the population, with four sizes of community (SOC) representing the remaining 70% of the Population.[4]

NOTE: The four size-of-community categories within the STOC classification are not equivalent to the four size-of-community categories within the SOC classification since the latter do not have the three extreme types of community extracted from them.

---

[4]The year 02 size-of-community categories are based upon the population of the community in which the school being assessed is located. The year 01 size-of-community categories were based upon standard metropolitan statistical areas (SMSA), therefore every individual in an SMSA defined as a particular SOC category was classified as belonging to that SOC. The four sizes of community are Big Cities, Urban Fringes, Medium-Size Cities, and Smaller Places.

The definitions of the extreme types of community were derived from an "occupation question" for the school (see Exhibit 6-1). One extreme group was selected in each of the three directions indicated by an exploratory analysis: (1) schools in rural areas where a high proportion of the people were professionals or factory workers, (2) city schools where a high proportion of parents were either not regularly employed or were on welfare and a low proportion were professional or managerial, and (3) near-city and city schools where a high proportion of parents were professional or managerial and a low proportion were factory or farm workers, not regularly employed or on welfare.

Exhibit 6-1

In-School Occupation Categories

Principal's questionnaire categories	Code
Professional or managerial personnel	A
Sales, clerical, technical, or skilled workers	B
Factory or other blue collar workers	C
Farm workers	D
Not regularly employed	E
On welfare	F

In year 02, there were no extreme type-of-community categories defined for individuals out of school since we did not obtain occupation data from Adults and 17-year-olds not in school.

Smaller extreme TOC (type of community) groups (less than 10%) would have been even more representative of the respective extreme TOC's; larger extreme TOC groups would have had more reliably determined percentages of success. The sizes of the three extreme groups--approximately 10% of all individuals assessed - were chosen as a compromise between more representative extremeness and greater reliability. In the definitions

below, the capital letters represent the codes for the categories in Exhibit 6-1, and the percentages refer to the year 02 assessment.

The four size-of-community (SOC) categories [before the three extreme types of community (TOC) have been extracted from them] are as defined as follows:

SOC-1, Big City. This category comprises 21.16% of the total sample and represents those individuals attending schools or living in a community within the city limits of a city greater than 200,000 (compare to STOC-4).

SOC-2, Urban Fringe. This category comprises 22.68% of the total sample and represents those individuals attending schools or living in the metropolitan areas served by a city with a population greater than 200,000 but outside the city limits (compare to STOC-5).

SOC-3, Medium Size City. This category comprises 19.07% of the total sample and represents those individuals attending schools or living in a city with a population between 25,000 and 200,000 (compare to STOC-6).

SOC-4, Smaller Places. This category comprises 37.09% of the total sample and represents those individuals attending schools or living in a community with a population less than 25,000 (compare to STOC-7).

The seven size-and-type of community (STOC) categories are defined as follows:

STOC-1, Extreme Rural. This category comprises 9.15% of the total sample and represents individuals attending schools in a community having a population less than 3,500. They are among those ranked highest on the rural index D-(C+2A). The communities comprising the Extreme Rural category were located within the following three SOC[5] categories:

---

[5]The year 02 TOC categories were extracted from the year 01 SOC categories.

Urban Fringe: 0.18%[6] --within SMSA[7] counties containing a city with a population greater than 200,000 but outside the city limits.

Medium Size City: 0.16%--within all other SMSA counties not containing a city with a population greater than 200,000 and other non-SMSA counties containing a city with a population between 25,000 and 50,000.

Smaller Places: 8.80%--within all other non-SMSA counties not included in the Medium-Size City Category.

STOC-2, Extreme Inner City. This category comprises 7.25% of the total sample and represents individuals attending schools in a community within the city limits or residential area served by a city with a population greater than 150,000. They are among those ranked highest on the inner-city index E+F-A. The communities comprising the Extreme Inner City category were located within the following three SOC categories:

Big City: 6.45%--within the city limits of a city with a population greater than 200,000.

Urban Fringe: 0.17%--within SMSA counties containing a city with a population greater than 200,000 but outside the city limits.

Medium Size City: 0.64%--within other SMSA counties not included above.

STOC-3, Extreme Affluent Suburb. This category comprises 12.15% of the total sample and represents individuals attending schools in a community within the city limits or residential area served by a city with a population greater than 150,000.

---

[6]These are percents of the total sample and, within each STOC, add to the percent representing the STOC.

[7]SMSA--Standard Metropolitan Statistical Areas.* An economic and social unit which is metropolitan in character and contains at least: (a) One central city with 50,000 inhabitants or more,or (b) two cities having contiguous boundaries with a combined population of at least 50,000. The smaller city must have a population of at least 15,000. The SMSA includes the county in which the central city is located, and adjacent counties that are found to be metropolitan in character and economically and socially integrated with the county of the central city.

*A few states have slightly different definitions for SMSA.

They are among those ranked highest on the suburb index, A-(C+D+E+F). The communities comprising the Extreme Affluent Suburb category were located within the following three SOC categories:

Big City: 6.30%--within the city limits of a city with a population greater than 200,000.

Urban Fringe: 4.98%--within SMSA counties containing a city with a population greater than 200,000 but outside the city limits.

Medium Size City: 0.87%--within other SMSA counties not included in the above.

STOC-4, Inner City Fringe. This category comprises 8.41% of the total sample and replaces the Big City SOC category. It represents those individuals attending schools in a community within the city limits of a city greater than 200,000 not categorized by STOC-2 or STOC-3 as Extreme Inner City or Extreme Affluent Suburb, respectively.

STOC-5, Suburban Fringe. This category comprises 17.35% of the total sample and replaces the Urban Fringe SOC category. It represents those individuals attending schools in the metropolitan area served by a city with a population greater than 200,000 but outside the city limits not categorized by STOC-1, STOC-2, or STOC-3 as Extreme Rural, Extreme Inner City, or Extreme Affluent Suburb, respectively.

STOC-6, Medium City. This category comprises 17.40% of the total sample and replaces the Medium-Size-City SOC category. It represents those individuals attending schools in a city with a population between 25,000 and 200,000 not categorized by STOC-1, STOC-2, or STOC-3 as Extreme Rural, Extreme Inner City, or Extreme Affluent Suburb, respectively.

STOC-7, Small City. This category comprises 28.29% of the total sample and replaces the Smaller Places SOC category. It represents those individuals attending schools in a community with a population less than 25,000 not categorized by STOC-1 as Extreme Rural.

Parental Education

This characteristic refers to the highest educational level attained by at least one parent of the respondent. This information was provided by the respondent. For the purpose of definition, high school refers to grade 9 through grade 12.

No High School. Neither parent has any formal education beyond the eighth grade.

Some High School. At least one parent has some formal education beyond the eighth grade, but neither parent has graduated from high school.

Graduated from High School. At least one parent has graduated from high school, but neither parent has any formal education beyond high school.

Post High School. At least one parent has some formal education beyond high school which may include business, professional, or trade school training as well as college or university training.

CHAPTER 4

NATIONAL ASSESSMENT SAMPLES

Overview of National Assessment Samples

4-1. The focus of the National Assessment of Educational Progress (NAEP) is on obtaining information about the proportions of children, teenagers, and young adults in the nation who respond in alternative ways to exercises in the subject areas assessed by NAEP--not only at one point in time, but at various times so that it will be possible to determine what changes in knowledges, understandings, skills, and attitudes are occurring. This information could be obtained by assessing the entire population or by assessing at each point in time a sample carefully selected from the population so that it represents the entire population. Complete enumerations, in which the entire population would be assessed, are much more expensive to conduct than assessing a sample selected from the population. Besides the factor of cost, other factors which favor the assessment of a sample over the assessment of a population include the opportunity to collect the data over a shorter period of time, which allows earlier reporting of the data collected, and the opportunity to use a smaller, more highly trained and closely supervised field administration staff, resulting in more accurate and better quality data collection. Factors such as these resulted in a decision by National Assessment planners to obtain data from representative samples selected from the populations in which we are interested rather than from the entire populations.

Four age groups were selected to represent the populations of children, teenagers, and young adults: 9-year-olds, 13-year-olds, 17-year-olds, and young adults 26-35 years of age. In addition to collecting data at these four age levels, we are interested in obtaining similar information for certain subpopulations of the age populations. The subpopulations are defined by region of the country and size of the community in which people live, color (Blacks and Whites) and sex. The definitions of some of these categorizations are presented in chapter 3; others are defined later in this chapter.

The objective of reporting results for these population groups requires that the sample be scientifically selected so that these populations are adequately represented and so that it is possible to draw inferences about these populations from

the data collected on the sample. To insure that valid inferences could be drawn, each person in the populations to be sampled was given a known chance (probability) of being included in our samples.

Details of National Assessment Samples

4.2. The Interrelationship of Sample Design to Other Aspects of National Assessment

National Assessment is viewed by sampling statisticians as a sample survey. As such, the planning of National Assessment has many features in common with the planning of all other sample surveys. A list of overall survey objectives must be prepared and all aspects of the survey design must be kept consistent with these objectives. The population or populations to be observed or measured must be defined. A method for selecting a sample of members from this population must be developed. Decisions must be made about the data to be collected. Methods of measuring the population members in the sample or collecting the data must be devised and the field work must be organized. A plan for the summarization and analysis of data collected must be prepared. The degree of precision for the principal estimates must be given consideration. The total cost of the survey must be kept within reasonable bounds.

The sample design is the method of selecting the members of the population which are to be measured. These selected members of the population are called the sample. The method of selecting a sample from a human population can greatly influence the cost of locating and measuring the members of the sample. Alternative methods of measurement can also vary greatly in cost and thus limit the size of the sample if the total budget is limited. The precision of estimates based on the sample data will be influenced both by the sample design and by the total size of the sample.

It soon becomes apparent that no single feature of a survey design may be developed independently of the other features. The sample design, in particular, is closely associated to all features of the survey design. In practice, these interrelationships may be explored by planning several alternative survey designs. Some of the alternatives may then be ruled unfeasible because of excessive cost or nonadherence to overall survey objectives. Comparisons between the remaining

planned methods may then be made on the basis of cost efficiency, expected precision of estimates, control of bias in the measurement and estimations processes, or general practicability. Pilot studies at this stage of the planning process may be employed to resolve the more difficult problems.

The general details of sample survey planning discussed above apply to the planning of National Assessment. Some of the specific details and results of this planning process as they relate to the sample design are discussed in the remainder of this chapter.

4.3. Statement of General Objectives

For a discussion of the general objectives of National Assessment, the reader is referred to chapter 1. A few of these general objectives are stated here because they directly affect the entire survey plan, and, in particular, the sample design.

1. The long range emphasis in National Assessment is to be an assessment of progress in education.

2. Results of National Assessment should be understandable to the general public.

3. Results from National Assessment are to be used to describe the performance of broad population groups on specific exercises within well defined educational objective areas.

4. New and different methods of collecting data were to be tried.

5. No individual participant in the National Assessment survey should be required to give more than one hour of his time.

In terms of the general survey design, objective (1) above required that the study be repeated so that changes over time may be evaluated. The first complete cycle of National Assessment must then serve two purposes: (1) to describe the present status of the outcomes of the educational process; and (2) to establish a level or "benchmark" for future comparisons.

Objective (2) most directly affected the development of reporting procedures used in National Assessment. Objective (3) required that the results of National Assessment be reported by individual exercise. Emphasis on any individual person's

test score in National Assessment is completely lacking. The plan for summarization of collected data is to look at specific exercises and summarize the responses of all individuals within certain population groups. The objective of reporting results by population groups was a principal constraint on the sample design, since all of these populations were to be represented in the sample.

Objective (4) most seriously affected the cost structure of the field operations. This influence on cost subsequently affected the choice of an efficient sample design.

Objective (5) meant that each individual participating in National Assessment would only participate in a few of the total number of exercises assembled in a package. In terms of sample design requirements, this objective meant that the sample design should insure a representative probability sample for each package of exercises.

4.4 The Definitions of Population and Subpopulations

At an early stage of National Assessment planning, four age groups were selected as the target populations for National Assessment:

1. 9-year olds
2. 13-year olds
3. 17-year olds (defined as between 16 1/2 and 17 1/2), and
4. Young adults 26 to 35 years of age.

Definition of the target population by age rather than year or grade in school for the school age populations is one of the features of National Assessment that distinguishes it from most other educational surveys and offers a particular challenge in sample design and the organization of field procedures.

The target population in each of these four age groups was limited to persons residing within the 50 states and the District of Columbia. Certain persons who live in institutions or who are handicapped were assumed to be excluded from the target populations.

Initial plans were developed to categorize the population within each age group into subpopulations based on four characteristics :

1. Region of the country
2. Type of community
3. Sex

The four regions which divide National Assessment subpopulations are:

1. Northeast
2. Southeast
3. Central
4. West

The states belonging to each of the four National Assessment region subpopulations are shown in chapter 3.

Another classification of the general population into subpopulations that was considered important by National Assessment planners was based upon community characteristics. Four categories were considered in the planning stages:

1. Large cities (above 180,000 population),
2. Urban fringe (communities adjacent to the large cities),
3. Middle size cities (25,000 to 180,000), and
4. Small town-rural (below 25,000).

4.5 The Data to be Collected - Subject Matter Areas

Ten subject matter areas were included in the long range plans for National Assessment. These areas are Art, Career and Occupational Development, Citizenship, Literature, Mathematics, Music, Reading, Science, Social Studies, and Writing.

Educational objectives were developed within each area and then specific exercises were developed within each area to focus on each objective.

Due to the large number of subject matter areas, objectives within area, and exercises within objective, it became apparent that a complete cycle of National Assessment covering all ten subject matter areas would be an extremely large project. Also, since one of the primary objectives of National Assessment was to assess change or progress over time, the team assembled to accomplish this task would have to be disbanded after the first cycle and reorganized five years later to obtain data for making comparisons. Such a plan, although feasible, had serious drawbacks. A continuous operation cycling plan involving two or more subject matter areas each year was developed. This plan allowed for including subject matter areas in the cycle about every five years.

4.6 Methods of Data Collection - Packaging and Administration of National Assessment Exercises.

One of the principal objectives of National Assessment was to try new methods of collecting data on educational outcomes. These methods were not to be limited by convenience factors for scoring to simple paper and pencil exercises. Each exercise was to be designed toward a specific objective within a single subject matter area; for example, a respondent should not be hampered on a Science exercise by his inability to read. As a result of this general objective, National Assessment exercises have many unique features.

Most exercises are read to the respondents. A standard taped voice presentation is used for group sessions conducted in schools. Other exercises require some interviewer-respondent interaction; these must be administered by specially-trained National Assessment exercise administrators. Other exercises can only be administered with the aid of special equipment and on a closely-supervised individual basis; typical exercises of this type are found in Science. Some of the Citizenship exercises used in year 01 of National Assessment involved the observation of discussion groups; two specially trained exercise administrators were required to administer and observe performance on these exercises. Other exercises, such as in the Music area, will require the recording of a respondent's performance on magnetic tape.

One immediate consequence of the many unique features of National Assessment exercises was that specially trained personnel would be required to supervise and conduct the

administration of these exercises. The use of a centrally trained group of supervisors also had the advantage of maintaining uniformity of administration methods in all parts of the country.

A second principal objective of National Assessment that influenced the methods of data collection was the limitation of individual participation to one hour or less. This did not limit the number of exercises since each individual respondent is not required to attempt all the exercises. Instead, all the exercises prepared for assessment of one age group during one year of National Assessment are sorted into sets of exercises and assembled in what is called a "package." Ten or more packages may be required for each age group. Each package may contain exercises from more than one subject matter area. A variety of types of exercises may be included. To gain efficiency through group administration in the school part of National Assessment, most of the exercise which could be administered to several persons assembled in a group were packaged in one set of packages, called group-administered packages. Those exercises which required individual administration were packaged in some limited number of individually-administered packages. The numbers of packages by age group and method of administration are illustrated for Year 02 in Exhibit 4-1.

Exhibit 4-1

Numbers of National Assessment Packages
by Age Group and Method of Administration for Year 02

	Method of Administration	
Age Level	Group Administered	Individually Administered
9-year olds	9	3
13-year olds	13	2
17-year olds	10	2
Young Adults	---	6

4.7. The Need for Probability Sampling

Most of the preceding discussion of National Assessment planning would be equally applicable for a complete enumeration survey of each of the target populations or for a sample survey of the target populations. The objectives of National of National Assessment might be met by a complete enumeration

type survey, but the cost would be unreasonably high. The decennial census of population is an example of a complete enumeration survey of the entire population. Since the four target age groups are well mixed with the general population, the effort required for a complete enumeration National Assessment would be similar to that required to conduct the decennial census.

A sample survey based on a probability sampling method allows researchers to collect data from a small sample of the population and to infer from that sample certain characteristics of the entire population. In particular, if one is interested in certain population-average values, totals, or ratios, these values may be estimated using data collected from the sample. In addition, if the sampling error of the estimate can also be estimated from the sample data, statements about the precision of the estimates can be made. A point to be stressed here is that sampling error can only show how estimates computed from the sample may differ from the corresponding population values that would be obtained in a complete enumeration survey using identical data collection procedures.

Other types of errors in the estimate, such as errors due to nonresponse, recording errors, processing errors, or others are not reflected in the sampling error. These kinds of errors are called non-sampling errors and could occur even in a complete enumeration survey. Because a sample survey involves a smaller work load, more attention can be given to supervision and training of personnel and this may in fact tend to reduce the non-sampling errors.

Other factors besides cost limitations and control of non-sampling errors which favor a sample survey in general over a complete enumeration survey are the opportunities to collect data over a short period of time and the opportunity to use more specialized techniques because of the smaller field force being trained and supervised. Both of these factors are important in the National Assessment survey.

Nonprobability sampling methods are sometimes used in conducting research projects. Samples are made up of volunteers or of "typical units"chosen on the basis of expert judgement Although such methods may be adequate for certain restricted purposes, they do not allow any valid inference about the total population. Because of the importance placed on describing entire populations and specified subpopulations in National Assessment, no serious consideration was ever given to non-probability sampling methods.

4.8. Sample Size

Guidelines were needed early in the planning stages on sample size of the number of respondents required for each exercise. These guidelines were to be based on a reasonable level of precision for key estimates. The stated long-range objectives of National Assessment involve the assessment of progress over time. It was decided to consider the sample size required to obtain sufficiently precise estimates to detect changes in the performance of a subpopulation on specific exercises.

Initial guidelines were based on a simple random sampling scheme. In practice, when other sampling schemes must be used, the sample size required to achieve the same precision will generally be greater by a factor called the design effect. Exhibit 4-2 shows the required sample sizes if one wishes to correctly decide whether a specified change has occurred over time or if no improvement occurred. The sample sizes in Exhibit 4-2 are computed so that decision of change versus no-change can be made correctly 9 out of 10 times.

Based on a graphical interpolation of Exhibit 4-2 a subpopulation "effective" sample size of 400 would be adequate to decide correctly in 9 out of 10 cases on the average if positive changes of the following types have or have not occurred:

1. A change of about .04 from .90 to .94
2. A change of about .06 from .10 to .16
3. A change of about .09 from .50 to .59

Exhibit 4-2

"Effective " Sample Sizes Required to Decide Correctly 9 out of 10 Times Whether Specified Positive Changes in a Proportion Have Occurred Over Time

Specified Change from Time 1 to Time 2	Time Level at Time 1		
	.10	.50	.90
+.025	---	---	1660
+.05	571	1311	355
+.10	164	325	---
+.15	81	142	---
+.20	50	78	---

Since some of the principal target subpopulations such as region and community type partition the U.S. population into

at least four groups, four times 400 or 1600 would be considered as the appropriate effective sample size on a national basis to detect real positive changes over time of the type shown above. If the design effect were larger than 1, a somewhat larger sample size would be required.

Consideration such as these led NAEP planners to consider sample sizes of 2000 to 2500 respondents per exercise. Since exercises were packaged in several packages for each age group, the total number of package responses required from an age group was 2000 times the number of packages. As a hypothetical example, if 10 packages were used for each age group, the total number of package responses planned would be 20,000 (2,000 x 10). For the four age groups, the total number of package responses required for one year's National Assessment survey would be 80,000 (20,000 x 4).

Actual figures on planned numbers of package responses differ from this hypothetical example because of varying numbers packages by age group and because of special adaptations of the sample design to the type of exercises contained in each package. Specific figures on planned sample sizes for year 02 of National Assessment are given in sections 4.11 and 4.16.

4.9. Sampling Frames

In order to select a probability sample, it is necessary to have a list of sampling units. Such a list is referred to as a sampling frame or frame. The ultimate objective of the sampling process is to select a sample of observational units. In this case, observational units are the persons in each age group who are eligible to participate in National Assessment. In some surveys, it is possible to list all the observational units and in such cases, the sampling units are the same as the observational units. In most surveys of human populations, it is not feasible to list all the eligible participants or observational units. Some other list of sampling units may be much more convenient; as an example, it is much easier to compile a list of school buildings than it is to compile a list of all the students attending classes at these school buildings. Similarly, it is much simpler to compile a list of all household addresses or locations in a county than it is to list all household members in the county. If such units as school buildings or households are used as an alternative to listing all observational units, it is necessary then to have a rule of association that identifies each observational unit with exactly one sampling unit. For example, a nine-year-old may be associated with the school at which he regularly attends classes. Most individuals can be associated with the household of which they are a member; some specific rules are needed to establish what constitutes membership in a household. A

few persons may be excluded from one or both of these two types of frames because they do not go to school or do not belong to any household.

In practice, further economies in listing may be achieved by associating households with certain identifiable land areas called area segments. The area segments may be associated with a larger area such as a city or county. Schools may also be associated with the city or county in which they are located.

Devices for constructing sampling frames such as discussed above allow samplers to select the sample observational units by first selecting a sample of primary sampling units which may be cities or counties. Further frame listing then is required only within the selected units. At each stage of sample selection, frames are constructed for the next stage of sampling only within the selected units. Eventually observational units are listed at the final stage of sampling but only for a very small proportion of the total population when compared with a list of all the observational units in the target population.

A basic issue in the planning stages of National Assessment was the choice of sampling frame. Several options considered included (1) a school frame, (2) a household frame, (3) a mixed household and school frame, and (4) frames based on other existing surveys.

All the listed options have some disadvantages. A school frame clearly cannot be used to identify the young adult target populations. A certain proportion of 9's, 13's, and 17's also may not be enrolled in school. Since most states have laws requiring enrollment in school up to some minimum age, most 9-year-olds and 13-year-olds may be expected to be in school. A lower proportion of 17-year-olds may be expected to be enrolled. Census estimates for the 1965 noninstitutional population show 99.3 percent of persons 7 - 9 years of age enrolled at the beginning of the school year (October), 99.4 percent of persons 10 - 13 years of age, and 93.2 percent of persons 14 - 17 years of age.[1] Estimates of percent of persons 16 1/2 to 17 1/2 enrolled vary by the time of this time of the year. For planning purposes, it was assumed that 75 to 80 percent of 17-year-olds are enrolled in school.

---

[1]U.S. Bureau of Census, Pocket Data Book, USA, 1969, p. 153.

Since persons in all four age groups could be identified with a household frame, this option was given consideration. In order to identify eligible respondents in each age group using a household sample, a roster of household occupants by age must be obtained for each household in the sample. Those persons falling in the defined target age populations are then asked to participate in National Assessment. This process is called household screening. Based on 1960 population estimates by age group and assuming a total sample in each age group of 20,000 persons, it would be necessary to screen 8 out of every 10,000 households to obtain an adequate number of young adults (26-35). In order to obtain the same number of 9-year-olds, it would be necessary to screen 50 out of every 10,000 households. A high proportion of the field survey costs associated with a household survey of this type are screening costs.

It was recognized that 9-, 13-, and 17-year-olds could be identified at much less cost using a school sampling frame. Certain other economies, such as group administration of certain packages, could be employed with the school frame to further reduce costs.

A shortcoming of the school sampling frame was the possible loss of a large portion of the sample due to lack of cooperation from state or district school administration officials. This factor was recognized by the NAEP planners and survey plans included allowances for special efforts to obtain the necessary cooperation whenever at all possible.

The final choice of sampling frames for National Assessment involved the use of a school frame for 9-, 13-, and 17-year-olds enrolled in school and a household frame for young adults. No effort was planned to locate 9-, and 13-year-olds not enrolled in school. Seventeen-year-olds not enrolled in school were to be screened for and identified along with the young adults using the household frame, but the number of households screened was to be limited to the number required for the young adult sample.

In order to increase the number of out-of-school 17-year-olds identified in the household screening process, the definition of out-of-school 17-year-olds was expanded to include all persons 16 1/2 to 18 1/2 years old who were not enrolled in school when they were 16 1/2 to 17 1/2 years old. This definition assumed that the performance of an individual on a set of exercises would not change over the period of a year if that individual were not enrolled in school. Use of this definition was expected to double the number of out-of-school 17-year-olds identified in the household screening process.

Since the total number of out-of-school 17-year-olds identified in the household screening process would still be much smaller than a proportional allocation would demand, other methods of identifying such persons were considered. The use of other national surveys, such as the Bureau of Census, Current Population Survey (CPS), as a screening device was considered. Special procedures are required by the Bureau of Census to obtain permission from individuals in the CPS sample to permit the release of their names to an outside survey organization. Due to the probably poor success of these procedures, it was decided not to attempt the use of the CPS survey as an additional frame.

In year 01 of National Assessment, no additional supplemental sampling frames for out-of-school 17-year-olds were used. As a result, about 500 such individuals were contacted in the sample. Year 02 plans called for some developmental work to develop and test alternative sampling frames for out-of-school 17-year-olds. This work is discussed fully in Section 4.19.

4.10. Target Populations for the School Sample

As discussed in Section 4.9, the school sample is aimed at three of the four age levels: age 9, age 13, and age 17. No other means of sampling is used for ages 9 and 13. Further attempts to identify and sample 17-year-olds who are not enrolled in school are made through the household sample which is also used primarily for adults and through special supplementary sampling frames. Both of these other sampling methods for age 17 are discussed in Section 4.15 to 4.19.

The field operation for year 02 of National Assessment was scheduled so that each of the age levels was assessed during a period of approximately two months with the assessment of age 13 beginning in October, age 9 in January, and age 17 in March. Students in sample schools were considered eligible to be selected into the National Assessment sample if their birthdates fell within certain specified ranges. Thus birth date-eligibility requirements were based on calendar years for 9- and 13-year olds and a special year-long range not coinciding with a calendar year for 17-year-olds. The same criterion for eligible birth date range was used throughout the survey period. Exhibit 4-3 shows the date of each survey and the eligible birth dates. Exhibit 4-4 is calculated from Exhibit 4-3 and shows the extreme in the age ranges that could occur from application of the particular eligibility criteria. An average age for each age group is also shown. Based on these calculations, the average of 9-year-olds was 9 years and 7 months, the average of 13-year-olds was 13 years and 4 1/2 months, and the average age of 17-year-olds was 17 and 0 months.

Exhibit 4-3

Specific Definitions for Year 02
School Sample Target Populations

Age Group	Survey Period	Eligible Birth Date
9	1/4/71 to 2/26/71	Calendar year 1961
13	10/12/70 to 12/11/70	Calendar year 1957
17	3/8/71 to 4/30/71	10/1/53 to 9/30/54

Exhibit 4-4

Range of Eligible Ages Based on The
Survey Period and Eligibility Rules

Age Group	Possible Range*	Average Age**
9	9 years to 10 years, 2 months	9 years, 7 months
13	12 years, 9 1/2 months to 13 years, 11 1/2 months	13 years, 4 1/2 months
17	16 years, 5 1/2 months to 17 years, 7 months	17 years

4.11. Overall Sample Size

Planned sample sizes for each National Assessment package were determined according to guidelines discussed in section 4.8. The actual numbers also were ultimately determined by considering the number of primary sampling units and the planned sample size within each primary sampling unit. A sample size of 2,160 was planned for each individually administered package. The precision of estimates based on group administered packages was expected to suffer from a higher design effect due to more clustering of samples with this method of administration; therefore, a sample size of 2,592 per package was planned for the group administered packages. Exhibit 4-5 summarizes total sample sizes planned by age group, by package, and by package types. The total planned sample size for 9-year-olds was 29,808; for 13-year-olds, 38,016; and for 17-year-olds, 30,240. The total planned sample size for the school sample for all three age groups was 98,064. Exhibit 4-6 summerizes the average

sample sizes by age group and by package types obtained in year 02.

Exhibit 4-5

Planned Sizes by Age Groups
in Year 02 of National Assessment

	Age Group		
	9-year-olds	13-year-olds	17-year-olds
Number of individually administered packages	3	2	2
Sample size for each package	2,160	2,160	2,160
Total sample for individually administered packages	6,480	4,320	4,320
Number of group administered packages	9	13	10
Sample size for each package	2,592	2,592	2,592
Total sample for group administered packages	23,328	33,696	25,920
Total planned sample	29,808	38,016	30,240

4.12 Sample Sizes Within a Primary Sampling Unit

On a per package basis, sample sizes of 20 per package for individually administered packages and 24 per package for group administered packages were planned for each regular two-week primary sampling unit (PSU). Each group administered package was to be administered to two group sessions of 12 students each. This general plan is illustrated in Exhibit 4-7. The total planned sample in a regular PSU was 276 for 9-year-olds, 352 for 13-year-olds and 280 for 17-year-olds. The sample for each age varied according to the number and type of packages. The total for all age groups was 908.

Exhibit 4-6

Average Size Per Package by Age Groups
In Year 02 of National Assessment

	Age Group		
	9-year-olds	13-year-olds	17-year-olds
Numbers of indiviually administered packages	3	2	2
Average sample size per package	2203	2199	2135
Total sample for individually administered packages	6609	4398	4270
Number of group administered packages	9	13	10
Average sample size per package	2579	2552	2366
Total sample for group administered packages	23211	33186	23660
Average sample size per package	2579	2552	2366
Total sample for group administered packages	23211	33186	23660
Total sample size	29820	37584	27930

In a few primary sampling units, designated as one-week primary sampling units, sample sizes of 10 per individually administered package and 12 per group administered packages were planned. The details of this plan are not tabled, but the bottom line of Exhibit 4-7 shows the total planned samples by age group for a primary sampling unit.

In terms of the field procedures required to complete assessment in a single PSU for a single age group, the two-week PSU's were scheduled for a two-week visit by a specially trained

National Assessment District Supervisor. During these two weeks, the District Supervisor assisted by two or three locally hired exercise administrators, completed the selection of the sample students and conducted all the necessary individual and group package administration sessions. The one-week PSUs were handled in a similar manner but only one week was allowed for completing assessment.

Exhibit 4-7

Planned Sample Size by Age Group
For One Primary Sampling Unit

	Age Group		
	9-year-olds	13-year-olds	17-year-olds
Number of individually administered packages	3	2	2
Sample size for each package	20	20	20
Total sample for individually administered packages	60	40	40
Number of group administered packages	9	13	10
Sample size for each package	24	24	24
Total sample for group administered packages	216	312	240
Total planned sample for one Regular or two-week PSU	276	352	280
Total planned sample size for one-week PSU	138	176	140

4.13. General Structure of the Multi-Stage Sample Design

One term that can be used to describe National Assessment sampling is the term multi-stage. This term means that the sample was selected in stage. The advantages of multi-stage designs with respect to sampling frame development were pointed out in Section 4.9. Multi-stage designs can also be used to concentrate or cluster the sample and thus reduce field costs. In order to discuss multi-stage designs, it is necessary to speak about several types of sampling units; namely, primary or first-stage sampling units. Secondary or second-stage sampling units, third-stage sampling units, and so on.

Four stages of sampling may be identified in the National Assessment school sample design. The primary sampling units were geographic land areas consisting of one or more whole counties. The sample selected consisted of 116 primary sampling selected with probabilities proportional to an estimate of size. The estimated size measure was the number of 17-year-olds in each primary sampling unit. The principle of selecting sampling units with probabilities proportional to their size or their estimated sizes was used at all stages of sampling.

The same primary sampling units were used for all three target age group samples. Secondary sampling units or second-stage units were formed within primary sampling units by grouping schools by zip code areas. In most cases several zip code areas were used to form one secondary sampling unit. Secondary sampling units of this type were employed only in the highly populated primary sampling units; in very small primary sampling units, this stage of sampling was not necessary. It was hoped that this type of clustering, produced by forming secondary sampling units within the large PSUs would simplify field procedures and ease the inconveniences to the local school systems by reducing the number of different administrative units involved. Secondary sampling units were also used as means of insuring that some relatively high socio-economic areas and some relatively low socio-economic areas were included in the sample from each primary sampling unit. In most cases, two secondary sampling units were selected per primary sampling unit. Procedures for selecting secondary sampling units did not guarantee that the same secondary sampling units would be used for all three levels.

The third-stage sampling units were schools. A probability sample of schools was selected independly for each age group from the schools in the secondary sampling unit selected

for the age group. The number of schools to be selected in each second stage sampling unit was not fixed, but was determined by the approximate number of students in the eligible age level attending each school. In other words, if the schools had very low enrollment more schools were required in the sample in order to obtain the prescribed number of students for participation in National Assessment. A general guideline imposed on the sample design at this point was that at least two schools be selected from each secondary sampling unit so that no single school would be required to supply a large proportion of the respondents within the PSU.

The final stage of sampling was the selection of a probability sample of students from the eligible age group at each sample school. The fourth stage units were then the students who were eligible to participate by the birth date rules. As students were selected it was also necessary to assign a specific package to each student. Special procedures were developed for selecting sample students and assigning packages which allowed some packages to be administered in group sessions of 12 students per session and other packages to be administered on an individual basis. These procedures also insured that the students participating in any particular package were a probability subsample of all the students participating in National Assessment.

4.14. Special Requirement Imposed on the School Sample Design

One change that occurred between year 01 and year 02 of National Assessment was the addition of the requirement that all states be included in the sample. This requirement did not in any way imply that the sample in each state should be adequate for reliable state estimates.

This requirement meant that if at least one two-week primary sampling unit was selected from each state, certain states with low total populations would have a larger proportion of the total sample, thus would be prescribed by allocation proportional to population. Sampling schemes that deviate greatly from a proportional allocation will generally produce estimates for the total population with a lowered precision per dollar. The use of one-week PSUs in these small states helped to alleviate the problem of disproportionate allocation.

4.15 Target Populations for the Household Sample

As discussed in section 4.9, the household sample was aimed at two of the four age groups: 17-year-olds not enrolled

in school and young adults 26 to 35 years of age. No other means of sampling was used for the young adult population. Approximately 90 to 91 percent of the 17-year population were enrolled in elementary and secondary schools. These eligibles were identified and assessed through the school sample. The household sample and the supplementary frames were used to sample the remaining 17-year-olds who were not enrolled in school.

The field operation for year 02 of National Assessment was scheduled so that the young adults and out-of-school 17-year-olds were assessed during a period of approximately five months beginning in March, 1971 and ending in July, 1971. The majority of eligibles sampled through the household survey were assessed during April, May, June and July. Most of the out-of-school 17-year-olds sampled through special frames were assessed during June and July. Sample persons were considered eligible if their birth dates fell within certain specified ranges. In addition to young adults and out-of-school 17-year-olds, an additional age group was assessed. This group was defined as 18-year-olds who were not enrolled in school when they were 17-year-olds. While this group was not part of the population of 17-year-olds, it was assumed that they would perform essentially the same way on National Assessment exercises as the out-of-school 17-year olds. Exhibit 4-8 shows the survey period and the eligible birth dates for the target populations. The survey period coincides approximately with the survey period for the in-school 17-year-old assessment Exhibit 4-3. The eligible birth dates for out-of-school 17-year-olds (Group A) are identical with the eligible dates for 17-year-olds assessed in the school sample. The eligible birth dates for 18-year-olds (Group B) are exactly one year earlier than for 17-year-olds. Exhibit 4-9 is calculated from the Exhibit 4-8 information and shows the extremes in actual ages which could occur using the eligibility criteria. An average age for each age group is also shown. The average age for young adults was 31 years, 4 1/2 months, based on these calculations. For the 17- and 18-year-olds, the average ages were 17 years, 2 1/2 months and 18 years, 2 1/2 months, respectively. However, both the 17- and the 18-year-olds include only individuals who were not enrolled in school when they were approximately 17 years of age (see footnotes to Table 4-9).

4.16. Overall Sample Size

Planned sample sizes for the year 02 National Assessment household survey are shown in Table 4-10. The planned sample size for each package is smaller than 2,000 to 2,500 determined

in early project planning and discussed earlier in Secion 4.8. The reasons for the smaller planned sample size necessitate reference to two previous surveys - the year 01 National Assessment household survey and the subsequent quality check survey. The year 01 household survey results did not approach the completeness or quality of the year 01 in-school assessment. In summary, a household screening rate of 77% and an individual completion rate of 57% were achieved. Thus, the overall rate was approximately 44%. The quality check survey results indicated that more experienced interviewers, experienced constant supervision and monetary incentives to respondents could raise the overall completion rate for young adults to over 70% (91-1/2 percent household screening completion and, with the incentives, 80 percent completion rate for eligibles in screened households). To test these results, the year 02 planned sample sizes were set at approximately one-half those which might be regarded as full-scale operational sample sizes for National Assessment. The household sample was limited to the number of households required to produce approximately 1,000 responses for each of the six young adult packages. Out-of-school 17-year-olds were screened and assessed using the household frame, but the bulk of the out-of-school 17-year-old respondents were assessed from supplemental list frame surveys. Three supplemental frames were sampled in year 02 of National Assessment. About two-thirds of the sampling was from lists of high school dropouts provided by a sample of schools. Enrollees at Job Corps and Neighborhood Youth Corps sites were sampled also; these made up approximately one-third of the supplemental list sampling for out-of-school 17-year-olds.

All packages were administered individually to young adult and out-of-school 17-year-old respondents in the household survey. In addition, each respondent was given the option of completing one, two, three or four assessment packages. It was decided to offer an incentive of $5.00 per package to the respondents agreeing to complete two or more packages, based upon the results of the year 01 household survey quality check.

For the supplemental frame surveys, all packages were administered individually to 17-year-old respondents from the school dropout list sample. These respondents were also given the option of completing up to four assessment packages per respondent with the same monetary incentives offered household survey respondents. Some of the packages were administered to groups of eligibles from the Job Corps and Neighborhood Youth Corps samples. Exhibit 4-11 shows the planned sample sizes for group and individual packages in all three of the supplemental frame surveys.

Exhibit 4-8

Specific Definitions for Year 02
Household Sample Target Populations

Age Group	Survey Period	Eligible Birth Dates
Young Adults 26 to 35	3/1/71 to 7/31/71	4/1/35 to 3/31/45
17 Year Olds		
Group A[1]	3/1/71 to 7/31/71	10/1/53 to 9/30/54
Group B[2]	3/1/71 to 7/31/71	10/1/52 to 9/30/53

[1]Not enrolled in school 1/1/71 to 1/31/71

[2]Not enrolled in school 3/1/70 to 3/31/70

Exhibit 4-9

Range of Eligible Ages Based on Survey
Period and Elibibility Rules

Age Group	Possible Range	Average Age
Young Adults 26 to 35	25 years, 11 months to 36 years, 4 months	31 years, 1 1/2 months
17 Year Olds		
Group A[1]	16 years, 7 months to 17 years, 10 months	17 years, 2 1/2 months
Group B[2]	17 years, 7 months to 18 years, 10 months	18 years, 2 1/2 months

[1]Not enrolled in school when 16 years, 3 months to 17 years, 4 months of age.

[2]Not enrolled in school when 16 years, 5 months to 17 years, 6 months of age.

Exhibit 4-10

Sample Sizes for Year 02 National
Assessment Household Survey, by Age Groups

Estimated Number of:	Age Group: Young Adults	Age Group: Out-of-School 17's
Households	6,925	6,925
Eligibles	2,547	76
Respondents	1,783	53
Completed packages	6,240	185
Number of packages	6	12
Planned Sample Size for each package	1,040	15.4
Altered Average Sample Size per Package	1,244	27

Each of the respondents in the Neighborhood Youth Corps and Job Corps samples were asked to complete two packages each - either two group packages or the two individually administered packages. An incentive of $10.00 per respondent was offered.

4.17. Sample Size with a Primary Sampling Unit

On a per-package basis, each young adult package was expected to be administered approximately 20 times in each primary sampling unit (PSU) in the household survey. It should be noted that the planned sample sizes per PSU are averages for the household sample rather than fixed numbers, as in the school sample. Exhibit 4-12 illustrates the average number of households, eligibles, respondents and completed packages for the year 02 household survey.

Within each PSU, a National Assessment Field Supervisor visited the PSU initially to list several sample segments and interview potential field interviewers. The Field Supervisor made additional periodic visits to list additional segments, assist the field interviewers with problems, conduct quality check work and give further training to the field interviewers. There were a total of 52 PSUs in the year 02 household sample.

Exhibit 4-11

Planned Sample Sizes for Year 02 National Assesment of Out-of-School 17-Year-Olds from Supplemental Frames

	Supplemental Frames		
	School Dropout Lists	Job Corps Sites	Neighborhood Youth Corps Sites
Number of individually administered packages	12	2	2
Sample size for each package	102	20	20
Total sample size for individually administered packages	1,225	40	40
Number of group administered packages	-	10	10
Sample size for each package	-	24	24
Total sample size for group administered packages	-	240	240
Total planned sample size:			
Completed packages	1,225	280	280
Number of respondents	350	140	140

Exhibit 4-12

Planned Average Sample Size by Age Group for One Primary Sampling Unit Year 02 National Assessment Household Survey

Estimated Number of:	Age Group: Young Adults	Age Group: Out-of-School 17's
Households	133	133
Eligibles	49.0	1.5
Respondents	34.3	1.0
Completed Packages	120	3.6
Number of Packages	6	12
Sample Size for each package	20	.30

Each PSU contained ten sample segments of approximately 13 occupied housing units each.

The supplemental frame samples were specially designed to survey out-of-school 17-year-olds only. The year 02 National Assessment school sample was subsampled as one of the supplemental samples. Lists of Job Corps sites and Neighborhood Youth Corps sites were also sampled as supplemental frames. The selection of samples from these supplemental frames is also discussed in later chapters. Exhibit 4-13 shows planned sampled sizes per PSU for the supplemental frames used in the year 02 National Assessment.

4.18. Structure of the Household Sample Design

The National Assessment household sample was designed as a multistage sample. The advantages of multistage designs were discussed in section 4.9.

The primary Sampling units were geographic land areas consisting of one or more counties. The sample consisted of 52 PSUs selected with probabilities proportional to a measure of size. The size measures used were in the 1970 census preliminary county populations.

The secondary sampling units were land area segments or clusters of housing units within the primary sampling units. The area segments were defined to contain an average of approximately 13.3 occupied housing units and 4.9 eligible young adults. The secondary sampling units within each PSU were stratified into five socio-economic status substrata and two sampling units were selected with equal probabilities from each substratum.

The third-stage sampling units were the occupied housing units within the secondary sampling units and the fourth-stage sampling units were eligible young adults and out-of-school 17-year-olds within the occupied housing units. There was no subsampling within the secondary units; that is, all eligibles in all occupied housing units of the secondary sampling units were designated as "in the sample."

4.19. Structure of the Supplemental Frame Sample Designs

Multistage sampling designs were also used to select the samples of out-of-school 17-year-olds from supplemental frames.

As mentioned in section 4.17, a subsample of sample PSUs and sample schools selected for the year 02 school sample were used as the basic sample for the school dropout sample. Thus, the primary sampling units, secondary sampling units, and third-stage sampling units used for the school dropout sample were those discussed in section 3 of this report. One-half of the year 02 PSUs and one-half of the schools in the sample for the 17-year-old in-school assessment in those PSUs were selected for the school dropout sample. The fourth stage sampling units were high school dropouts on lists furnished by the sample schools. The lists of dropouts contained some individuals not meeting the National Assessment eligibility criteria shown in Exhibit 4-8.

Two-stage sample designs were used for the Job Corps and Neighborhood Youth Corps supplemental frame samples, which were selected independently. The first-stage sampling units in both cases were Job Corps sites and Neighborhood Youth Corps Centers, identified on lists obtained from U.S. Department of Labor. Primary samples consisting of five Job Corps sites and five Neighborhood Youth Corps centers were selected using probabilities proportional to an estimated size measure. The measures of size used were the site and center capacities, in terms of number of enrollees.

The second-stage sampling units were enrollees at the Job Corps and Neighborhood Youth Corps PSUs who were eligible for National Assessment. Eligibles were identified from records available at the PSUs. Equal-sized samples of eligibles from each site were selected from these lists of eligibles with equal probabilities.

4.20. Special Requirements Imposed on the Household Sample Design

As mentioned in section 4.16, the year 02 National Assessment household sample size was approximately one-half what would be considered the full-scale National Assessment sample size. Problems occurred in the year 01 household survey which led to a decision to use a smaller sample in year 02 to test whether modifications in the field procedures would lead to more acceptable results. It was expected that the year 03 household sample would be expanded to the operational level of 2,000 to 2,500 responses for each young adult package, assuming that the response rates using the modified field procedures were acceptable. Thus, for reasons of economy, the primary sample used in year 02 was required to be designed as an expandable sample. That is, a larger primary sample might be required for the year 02 household sample and it would be more efficient, cost-wise, to have the larger year 03 primary sample of PSUs include the 52 PSUs selected for the year 02 primary sample. This requirement was met by selecting 208 PSUs with probabilities proportional to size and subsequently selecting 52 of the 208 PSUs with equal probabilities for the year 02 household survey primary sample. Additional PSUs could be selected from the 156 remaining PSUs for use in year 03 and the overall probabilities of selection for PSUs in the expanded sample would still be proportional to size.

Future plans included the possibility that the same primary sample might be used for several years. Thus, PSUs were required to contain sufficient respondents for several years of household assessment without assessing any respondents more than once. This requirement was met at the time the sampling frame was constructed by requiring that each PSU in the frame must include at least 20,000 population in 1970. Thus, counties with 1970 populations less than this minimum were combined with other counties prior to the primary sample selection. Another reason for establishing a minimum size for PSUs was the policy of oversampling the low-SES portion of the population. Since the oversampling was to be effected within sample PSUs for certain size of community strata, it was essential that all PSUs contain a large enough population to make oversampling feasible.

Unlike the National Assessment school sample, there was no requirement that all states and District of Columbia be included in the primary sample for the household survey. Consequently, no such restriction was imposed on the sample design.

Exhibit 4-13

Planned Sample Size for One Primary Sampling Unit for Year 02 National Assessment of Out-of-School 17-Year-Olds from Supplemental Frames

	Supplemental Frames		
	School Dropout Lists	Job Corps Sites	Neighborhood Youth Corps Site
Number of individually administered packages	12	2	2
Sample size for each package	1.76	4	4
Total sample size for individually administered packages	21.1	8	8
Number of group administered packages (per PSU)	-	4[1]	4[1]
Sample size for each package	-	12	12
Total sample size for group administered packages	-	48	48
Total planned sample for one PSU:			
Completed packages	21.1	56	56
Number of respondents	6.0	28	28

[1]Each of the ten group administered packages was assigned to two of the five Job Corps PSUs and to two of the five Neighborhood Youth Corps PSUs.

CHAPTER 5

ADMINISTERING THE ASSESSMENT

In this chapter, we discuss the procedures involved in the actual collecting of the data which, when processed, tell us what knowledges, understandings, skills, and attitudes various National Assessment groups[1] have attained in the 10 assessed subject areas.[2]

Most individuals who participate in the National Assessment project are enrolled in school at the time. This "in-school" sample includes all 9-year-olds, all 13-year-olds, and most 17-year-olds.[3]

In addition to assessing children and teen-agers enrolled in school, we assess two groups of individuals who are not enrolled in school. This "out-of-school" assessment includes: (1) adults aged 26-35 and (2) those 17-year-olds who have either dropped out of secondary school or who have completed secondary school at an early age and may or may not be enrolled in college or its equivalent.[4] The adult assessment provides information about the current knowledges, understandings, skills, and attitudes in the 10 assessment subject areas possessed by individuals most of whom have been away from formal education for a number of years. The out-of-school 17-year-olds are included in the assessment to make the results for 17-year-olds more representative of all 17-year-olds.

Since the details of the in-school assessment and the out-of-school assessment differ considerably, they are discussed in separate sections. The discussion is oriented to the year 02 assessment of Reading and Literature, but many of the procedures can be generalized to other assessment years.

---

[1]See chapter 3 for definitions of the National Assessment groups.

[2]See chapter 1 for a listing of the 10 subject areas.

[3]See chapter 3 for definitions of these age levels.

[4]See chapter 3 for a definition of the adult and the out-of-school 17-year-old populations.

The In-School Assessment

The National Assessment staff working on the in-school assessment was concerned with two major activities--placing the selected National Assessment exercises into administrative units called packages, or booklets, and administering the packages to 9-, 13-, and 17-year-old students in accordance with the in-school sampling plan (see chapter 4). The actual compilation of packages was done by the Research Triangle Institute (RTI) of Raleigh, North Carolina; the printing of the packages was done by the Measurement Research Center (MRC) of Iowa City, Iowa; and the field administration was conducted by RTI with a subcontract to MRC for the Central and Western regions. The National Assessment staff performed planning, coordinating, and monitoring functions for all administration processes.

At each age level, the selected exercises were packaged into 35 minute units. Most packages were designed to be administered to students in groups of 12. Some packages, how ever, were designed to be administered to individual students in an interview situation. For age 9, there were nine group-administered packages and three individual-administered packages; for age 13, there were 13 group packages and two individual packages; and for age 17, there were 10 group packages and two individual packages. The National Assessment packages were constructed according to these criteria:

1. Each package contained exercises from both Reading and Literature.

2. Each package contained exercises from the three levels of difficulty (easy, medium, and hard) and began with an "easy" exercise.

3. Some exercises could not appear in the same package due to similarity, keying, etc.

4. Exercises could not appear in the same package if they used the same, or similar, stimulus materials.

For each group package, a tape recording was produced. The tape first gave all introductory statements. Sample exercises were on the tape as well as printed in the booklet. For Literature, all exercise directions, written stimulus materials, exercise stems, and exercise foils were both on the tape and in the printed booklet. For Reading exercises which were dependent upon reading ability, only the exercise directions were on the tape. The tape assured uniformity of administration by different

exercise administrators and by the same exercise administrator from one administration to another. The tape also helped to prevent exercises for subject areas other than Reading from measuring reading ability instead of the subject area in question (e.g., Literature for year 02). The tape was paced to allow the proper amount of response time for each exercise. Students marked their answers directly into the packages.

In individual packages, directions were given by the interviewer to the student being interviewed. Poems or stories used as Literature stimuli were read to the student from a tape recording to assure uniformity in the stimulus materials. In most cases the interviewer wrote the student's oral remarks in the packages; however, the packages were handed to the students for certain exercises. The student responded directly into a tape recorder for one Literature exercise.

Thirty-five minutes of exercise time per package allowed the field staff to work generally within class periods of schools and to avoid using students for more than approximately 50 minutes. Introductory statements, completion of personal data questions, tape time not associated with exercises (in group packages) and miscellaneous administrative time generally required about 15 minutes per package administration.

Before the packages could be administered in selected schools, cooperation had to be obtained from school personnel and operational procedures had to be established between them and the National Assessment staff. Chief state school officers were informed that identification of all schools selected for the assessment within their respective states would be available upon request. Many states requested this information, and some then contacted superintendents to tell them schools from their districts had been selected for participation in the project. The first contact by National Assessment was a letter in late July, 1970, from the Staff Director. In addition to being notified of the selection of his schools in the sample, the superintendent or private school official was told he would receive descriptive materials from RTI or MRC, depending upon the geographical region. These materials informed the superintendent of the identity of all his schools selected for the year's assessment, gave the dates for all contact with him, his principals and other school personnel, described National Assessment, and informed him that a member of the field staff would telephone to arrange a meeting with him and the principals of schools selected for 9- and 13-year-old assessments. A proposed date for the meeting was included in the letter.

The Research Triangle Institute and the Measurement Research Center employed four Regional Supervisors (one for each of the geographical regions) and 29 District Supervisors. The District Supervisors were responsible for arranging the meetings with superintendents and principals, setting assessment dates, and conducting the actual assessment.

During late August, each District Supervisor called the superintendents in his district to answer questions about the project, confirm dates about selected schools, and arrange the meeting for the proposed date, or another mutually convenient date, in September or October. At the meeting the District Supervisor explained the purpose of National Assessment, described operational procedures for completion of the assessment, and determined weeks suitable for assessment in the area. In most areas, several school districts were to be involved so the District Supervisor suggested assessment weeks to be agreed upon by school personnel at the various meetings. Assessment of 13-year-olds was arranged for October, November, and December, 1970, and 9-year-old assessment was arranged for January and February, 1971.

Each principal appointed a coordinator, a staff member who had a flexible schedule, to prepare for the assessment. Prior to assessment week the coordinator arranged to have a Student Listing Form completed for each student born during the calendar year defining National Assessment eligibility (see chapter 3 for the definition of age levels). The District Supervisor used these cards to randomly determine the students to be assessed in the school (see chapter 4) and to obtain the birthdate, sex, grade, and color for each. When the District Supervisor arrived during assessment week the coordinator also arranged for space to conduct the group and individual administrations and worked with the District Supervisor to arrange a mutually convenient schedule in the school and to insure that students arrived on time for their scheduled administrations.

For 17-year-old assessment, the introductory meetings were held concurrent with the assessment of 9-year-olds, and the assessment was conducted during March and April, 1971.

After the sample was selected in a school, package administration was completed by the District Supervisor or by an Exercise Administrator who was hired locally and trained by the District Supervisor to professionally administer National Assessment packages. Exercise Administrators had various backgrounds including teaching, substitute teaching and survey research. Assessment time varied in schools depending upon the number of packages assigned, but it rarely lasted more than 2½ to 3 days.

As packages were administered, the District Supervisor checked all data coded on the package from Student Listing Forms for accuracy. When the District Supervisor completed assessment in a school, he instructed the coordinator to save these forms for six months in case they were needed for data verification. Since names were not associated with National Assessment packages, these cards which had been matched with a package by ID number were the only means of verifying lost or questionable data. After six months the cards were destroyed to assure the anonymity of students who participated in the assessment.

The District Supervisor mailed the packages to the scoring contractor where a quality check was performed to minimize errors in the data. Each entry from the Student Listing Form was both coded in arabic numerals and gridded for optical scanning. The codes and grids were checked against each other, and both were checked against acceptable ranges for the various entries. Errors which could not be reconciled by central staff were sent back to the District Supervisor to be checked against the Student Listing Forms.

Other quality checks included a small sample of schools from which information from the Student Listing Forms was obtained by central staff to be checked against packages returned from the field. This was to check the accuracy of transferring information recorded by the schools. Also, the central staff member discussed overall assessment procedures with the principal and coordinator in this sample to discover problems with field procedures and implementation. Each school which participated in the assessment received a thank-you letter and a questionnaire concerning contact procedures, preparation for assessment, assessment exercises, personal data questions, and the involvement during assessment week. The information was analyzed to discover problems in the field procedures which required modification for future assessments.

Out-of-School Assessment

The National Assessment staff working on the out-of-school assessment was concerned with placing the selected exercises into packages and administering the packages to the adults and out-of-school 17-year-olds in accordance with the out-of-school sampling plan (see chapter 4). The actual compilation and printing of the packages was done by the same contractors who performed these functions for the in-school assessment--Research Triangle Institute (RTI) and Measurement Research

Center (MRC), respectively. The field administration of the packages to the individuals in the out-of-school sample was handled exclusively by RTI and occured from March to August, 1971. As with the in-school assessment, the National Assessment staff performed planning, coordinating, and monitoring functions for all administration processes. Packages for the out-of-school assessment were constructed according to the same criteria as the in-school packages. In fact, the packages used in assessing the out-of-school 17-year-olds were the same as those used in assessing the in-school 17-year-olds.

Obtaining out-of-school individuals within the appropriate age ranges was more difficult than obtaining in-school students. The Field Supervisors personally identified and inspected the area segments to be sampled and compiled a list of all the sample housing units that fell within the area segment boundaries.

The Field Interviewers canvassed the sample housing units for eligible respondents in the two out-of-school age categories. They filled out a screening questionnaire on which they listed all household members and their exact birthdates. In this way, the screening process had the appearance of a simple survey. The interview was terminated if the screening questionnaire listed no eligible household members. Using a callback procedure for households where no one was at home, the Field Interviewers completed screening questionnaires on over 99% of the sample housing units.

The Field Interviewers then solicited cooperation in taking the assessment from those individuals who were deemed eligible on the basis of the screening questionnaires. The eligible individuals were offered a monetary incentive to complete more than one package: $10.00 for two packages, $15.00 for three packages, and $20.00 for a maximum of four packages completed. No monetary compensation was made for completing only one package. The four packages given to an adult or out-of-school 17-year-old were selected such that all packages within each age level were administered to approximately the same number of individuals and such that each package was administered in each ordinal position (first, second, third, or fourth) about the same number of times. Field interviewers were equipped with various National Assessment handouts in case a respondent wanted additional information about the project. If a cooperating eligible individual could not begin taking the assessment immediately, the Field Interviewer made an appointment to come back at a later time. The eligible individuals who refused to respond were noted by the Field Interviewer, and these cases

were reviewed by the Field Supervisor. In as many cases as possible, the Field Supervisor contacted those who refused in an attempt to convert the refusal to a participating respondent. This procedure produced successful results in many instances.

Although the out-of-school assessment is always a one-to-one situation (interviewer and respondent), the administration of the packages is kept comparable to the in-school mode of administration. The tape was not paced, however. The interviewer started and stopped the tape as necessary to allow the respondent a reasonable length of time to respond.

For the year 02 assessment of Reading and Literature, 75% of the eligible adults and 97% of the out-of-school 17-year-olds contacted consented to take at least one package. The average number of packages taken per respondents was 3.9.

The Field Supervisors were responsible for frequent contact with the Field Interviewers, both by personal visit and phone contact, to maintain a high degree of field quality and proficiency. The Field Supervisors revisited a sample of the housing units to validate the interviewers' work. RTI also used a mail-validation procedure to further safeguard field quality. Completed packages and screening forms were subject to quality checks upon being received at RTI, where field errors were recorded and brought to the attention of the field staff.

CHAPTER 6

SCORING OF EXERCISES

Each National Assessment exercise is scored in one of three ways: by a subject matter specialist (Professional scoring) by a well trained scorer (Semi-Professional) or by a machine.

Professional scoring is used for those items that demand difficult analyses, informed value decisions, or expertise in a particular subject area. Primarily, Professional scoring is necessary for the evaluation of essays, the subjective analysis of written material, or the evaluation of performance exercises--that is, exercises which require the respondent to perform in some prescribed manner best judged by an expert.

Exercises that are professionally scored may or may not require that the staff of subject area specialists be representative of various geographic regions and/or other considerations, but when geographic diversity is desirable, then Professional scoring is normally conducted in a work seminar at a central location. If, however, the only requirement is subject area expertise, then the staff is assembled from one locale and remain there for the scoring of exercises.

The semi-professional scoring is used for the scoring of items that contain short answers (such as one-word or title responses), brief essay responses, and listings or other comparable answers that are not as difficult or complex as those requiring professional scoring. There may also be a number of exercises that can be scored on a right-wrong basis depending upon the content of the response. In these instances the staff requirements are somewhat less than those required for professional scoring but follow the same pattern.

The third type of scoring, machine scoring, takes advantage of optical scanning devices that automatically score multiple-choice responses without the need for any individuals. This method provides a fast economical and practical way to get results into the computer system for analysis. Ultimately, all scores even the professional and semi-professional results are translated into terms that can be optically scanned by machine and fed into the computer system for faster analysis.

The educational requirements and experience for the professional and semi-professional staff are somewhat similar, but do vary in the degree of formal training and experience that are brought to the project:

Professional scoring staff requirements.

1. Professional scorers hold an advanced degree in the subject area under consideration. They either have a Ph.D. or are engaged in a Ph.D. program.

2. Professional scorers have actual experience in their major field and are able to bring to the program the results of this experience. If at all possible, this includes actual experience with the age groups under consideration in the National Assessment program.

3. All Professional scorers are approved by the National Assessment staff before they are included in the project.

Semi-professional scoring staff requirements.

1. Semi-professional scorers have a degree in the subject area field in which they are working.

2. Semi-professional scores have some actual experience in the field in their subject area and are able to bring the results of this experience to the project.

3. All semi-professional scorers meet the requirements of the specifications set forth by the National Assessment staff, and are approved by National Assessment before they are included in the scoring program.

Once a staff whether Professional or Semi-Professional, has been assembled for the scoring of the exercises for tne National Assessment program, a formal training program is conducted to acquaint the staff with the requirements of the project. This formal training program consists of three basic steps which are necessary for the proper scoring. These steps are as follows: (1) development of standards and guides for scoring, (2) the use of "live data" to verify the standards and guides and to provide actual scoring experience, (3) the resolution of discrepancies. The first step--the development of standard and guides--involves establishing the scoring criteria that are used and determining the type of information that is necessary for further reporting. These, then, act as the guides used by the scoring staff in the final analysis of the responses given by the participants.

The second step in the process, tne use of live data, is the actual training of the staff in scoring techniques. During this phase, the scorers work with actual tryout data from the field, and the results of their initial scorings are checked carefully to determine that scorers are, in fact, correctly utilizing the standards set forth for the exercises. In addition, it gives the scoring staff the opportunity to adjust or modify the standards and guides wherever necessary, based on tne results of their initial scoring of the live data.

The third and final step, the resolution of discrepancies, may take one of two forms. In the case of professional scoring, if the requirements for the subject area state that more than one reading or scoring of a given exercise is needed, then the resolution will be based on a second reading of the exercise and/or the comparative reading of a referee.

The second form is that of a quality check, which is a normal, ongoing procedure regardless of whether or not the exercises are professionally scored or semi-professionally scored. This consists of a check of the work of each of the scorers on a periodic basis to determine that there are no discrepancies and that they are, in fact, scoring the exercises according to the standards and guides that have been set forth. If any cases of discrepancy are noted, these can be corrected and the scoring adjusted so that the standards and guides are being followed as stated.

The final aspect of the scoring of the exercises before their inclusion in the analysis program is the scoring done by the optical scanning devices. The flexibility of the device allows for a complete range of consistency checks on the scoring as well as range checks on the numerical values assigned to the various parts of the exercises. This means there is an automatic screening or checking of each response and any exercise that has a response falling outside the ranges allowable, or that does not meet the consistency checks required by the standards, can be noted as a possible error to be corrected before its inclusion in the final process.

The end result of all of the above requirements and checks is data that does meet the standards of the National Assessment program for analysis purposes. The qualifications of the staff, the techniques that are set forth, and the checking of the end results all combine to provide the best possible scoring of the exercises, so that the data as reported is meaningful and constant.

CHAPTER 7

PROCESSING THE DATA

The ECS (Education Commission of the States) Data Processing Services Department (DPSD) was formed in May 1971 to provide data processing support for the National Assessment of Educational Progress project and other ECS projects with primary emphasis on the National Assessment Project. Two computers are used to process the data--one located in Denver and one located at the Princeton University Computer Center--both connected by high speed telephone lines.

Individual members of the DPSD staff are assigned the responsibility of Phase Coordinator for each assessment year with two or three other DPSD personnel as assistants. The Phase Coordinator is responsible for all data processing that takes place across all subject areas within an assessment year and is also a member of the NAEP subject area teams. In this way, he becomes familiar with all phases of the National Assessment Project--objectives development, exercise development, administration of the exercises, scoring specifications, and so on.

Student File

Approximately three months before the first MRC (Measurement Research Center) data tape is scheduled to arrive at the Princeton University Computer Center, DPSD personnel start designing and coding the NAEP student file. This file contains all information pertaining to each respondent. To ensure that the coding of the student file is correct, sample assessment packages are scored by MRC. This provides test data that is then sent to DPSD approximately one month before "live" data (the actual field administration results) is scheduled to arrive. The test data allows DPSD to find any program errors that might exist and make corrections before the live data is received. Magnetic data tapes arrive from MRC every six to eight weeks for each age level beginning with the exercise data for 13-year-olds followed by the 9-year-olds, in-school 17-year-olds, out-of-school 17-year-olds, and adults. Data for each age level is added onto the NAEP student file as DPSD receives it.

Work Files

The student file, which contains all the information collected for each respondent, is very large and therefore

cumbersome and costly to use in the many analyses required by NAEP. Therefore, smaller files, called work files, are created, each of which contains only the data required for the specific analysis to be performed. The programming staff, upon receipt of specifications from the analysis staff, writes computer programs to perform the desired computations. Summary files are created to store the results of these computations. The results and appendices reported in the summary and theme volumes can then be obtained from the summary file without having to redo the desired computations each time.

Appendix File

After the summary files are complete, the appendix file is created. This file is used to print a final report with a short description of the exercise, the national percentage of success, and group differences from the national percentage of success for each exercise, as well as standard errors and other documentation.

Final Reporting

At this point, DPSD is ready to begin the final reporting data analysis phase. The Research and Analysis staff sends specifications to DPSD which computes reports in many formats: statistical tables, data-distribution plots, and any other format requested by NAEP. A data base/data management system that is publicly available and widely used called NIPS is used to process the NAEP data. It is designed to retrieve large amounts of data quickly and efficiently.

CHAPTER 8

DESCRIBING NATIONAL ASSESSMENT DATA

Once an assessment has been completed, the responses scored, and the results analyzed, we become concerned with describing the results to educators and the concerned public in a manner that is understandable, meaningful, and useful. Since National Assessment data differ from typical test data, it is important that the reader have a clear understanding of our methods for describing data. National Assessment is not concerned with obtaining individual scores as are testing programs. We are concerned with the percentage of individuals in the nation as a whole and in certain groups (see chapter 3) who possess various knowledges, understandings, skills, and attitudes and with comparing the percentages to each other.

Percentage of Success

An exercise which assesses a knowledge, an understanding, or a skill has a correct/incorrect criterion; i.e., the exercise has a definite correct response. Essay exercises in the Writing assessment are exceptions to this rule, because they are assigned overall quality scores. An exercise which assesses an attitude--in many cases--has a desirable/undesirable criterion; i.e., the exercise has no definite correct response, but certain attitudes are considered to be more desirable than others. In some instances, it is not practicable to call an attitude either desirable or undesirable. We refer to correct and desirable responses, collectively, as acceptable, and we deem essays acceptable or unacceptable on the basis of their overall quality scores. The percentage of success for an exercise is the percentage of participants who made an acceptable response to that exercise. It is the number of respondents in a group who gave an acceptable response divided by the total number of respondents and multiplied by 100. Suppose, for example, there are 1500 respondents in a group; and, of these, 650 give an acceptable response. The percentage of success for the group would be:

# of acceptable responses	:	total # of responses	x 100 =	% of success
650	:	1500	x 100 =	43.3%

We can express a percentage of success for the nation as a whole or for any of the National Assessment groups (see chapter 6). For example, a 64% of success for 9-year-olds in the Northease region on a given exercise means that 64% of the 9-year-old respondents in the Northeast region gave an acceptable response to that exercise. Responses to those attitude exercises which cannot be classified as desirable or undesirable are reported as X% having this attitude and Y% having that attitude.

Derived Scores

Many exercises have several parts, each of which can be scored acceptable/unacceptable. For such an exercise, we often report the percentage of respondents who gave correct responses for none of the parts, for one part, for two parts, and so on. In the second case, we report the gave correct responses for none of the parts, for one part, for two parts, and so on. In the second case, we report the percentage of respondents who listed zero things, one thing, two things, and so on. We call these percents derived scores.

Comparisons among Groups

In most instances, we compare the percentage of success for a group with the percentage of success for the nation as a whole, and the number we obtain is called an effect. An effect is expressed as the percentage of success for a group minus the percentage of success for the nation as a whole. For example:

Northeast % of Success	-	National % of Success	=	Northeast Effect

A positive effect means that a larger percent of respondents in a group gave an acceptable response to an exercise than did so in the nation as a whole. For example, if 74% of 13-year-olds in the Northeast gave an acceptable response, but only 68% of 13-year-olds in the nation as a whole gave an acceptable response, the Northeast effect for 13-year-olds would be:

Northeast % of Success	-	National % of Success	=	Northeast Effect
74%	-	68%	=	+6%

A negative effect means that a smaller percent of respondents in a group gave an acceptable response to an exercise than

did so in the nation as a whole. For example, if 52% of 9-year-olds in the West gave an acceptable response, but 60% of 9-year-olds in the nation as a whole gave an acceptable response, the West effect for 9-year-olds would be

West % of Success	-	National % of Success	=	West Effect
52%	-	60%	=	-8%

Limitations

Many factors may affect an individual's ability to give acceptable responses to exercises in the subject areas assessed by National Assessment of Educational Progress. One reason for assessing these subject areas by characteristics such as size-and-type of community, region of the country, sex, color, and level of parental education is to describe what differences (if any) actually exist in the real world between members of various groups and the nation as a whole.

When we look at the results for groups within one characteristic at a time, however, we cannot say that a group effect is attributable solely to individuals being members of a group in question. A group--or the characteristic within which the group occurs--must not be construed as necessarily being the cause or even a cause of the effect associated with the group.

CHAPTER 9

OVERVIEW OF SELECTING EXERCISES TO BE REPORTED

The National Assessment policy for year 02 states that: (1) all exercises which are selected for reporting will be reproduced in the report exactly as they appeared in the assessment booklets and (2) at least 50% but less than 100% of the exercises in each of the two subject areas being reported will be selected for release. In order to implement part (2) of this policy, a procedure for the representative selection of exercises for release has been developed.

This chapter presents a brief non-technical description of the selection process; a more complete description of the process and its mathematical basis follows in chapter 10.

The primary purpose for developing a selection procedure was to insure that, although exercises would be selected randomly, they would nontheless be representative of the total pool of exercises available. With such a procedure we can be reasonably certain that a report provides coverage across objectives and across all population group differences to the extent that such coverage exists in the entire pool of exercises assessed.

It is critical in a report that includes only a portion of the exercise assessed that we have exercises which represent the entire spectrum of data. For example, there should be exercises for which males do much better than females, exercises which show no difference between males and females, and exercises for which the females do much better than the males. Our selection procedure enables us to achieve this kind of representative coverage for each group, i.e. males, females, NE, SE, etc.

After a computer randomly selects exercises for reporting, we fill whatever gaps remain by looking for exercises which will provide us with an example of the type of data which is not in the set of exercises to be released. For example, if none of the exercises that were selected for release represent a large female advantage, one or more exercises are selected specifically because they show a large female advantage. This systematic selection proved necessary for the year 02 Reading report but not for Literature.

Using systematic selection of 10 to 15 percent of the exercises, we are able to identify a set of exercises which are truly representative of the subject area being reported. In year 02, the selected exercises amounted to about 50% of all exercises.

CHAPTER 10

DETAILS OF SELECTING EXERCISES TO BE REPORTED

In this chapter, we examine the details of the mathematical procedures used to select exercises to be reported. We give simple examples for each step of the selection procedure so that the rationale (see chapter 9) and the procedure itself can be more clearly understood and duplicated if desired. Since the selection process differed slightly for Reading and Literature, we describe the general process first and then those procedures specific to each subject area.

Since the exercises not selected for release during the first assessment cycle are withheld to be released during the second assessment cycle, those exercises designated for immediate release and those withheld for later release should be equivalent in two ways. First, both sets of exercises must be equivalent in their coverage of objectives, themes, exercise formats and/or any other relevant characteristics. Second, both sets of exercises must be statistically equivalent; i.e., they must have similar representation across the entire spectrum of percentages of success. This latter requirement prevents currently reporting, for example, that girls can read charts better than boys (on the basis of the released exercises) and then reporting five years hence that boys can read charts better than girls (on the basis of the unreleased or withheld exercises). The same consideration applies to all reporting categories.

In order to insure the necessary equivalence National Assessment selects exercises in two steps. First, we group the exercises by their nonstatistical characteristics (objective, theme, format, etc.).[1] Then within each of these groupings, we attempt to achieve statistical equivalence by developing an index which reflects group differences and can be used to form sets of similar exercises.

---

[1] These groupings vary from one subject area to another, depending upon the nature of the exercises in a given subject area.

The groups[2] used to construct the selection index are:

Type of Community (TOC)

Extreme Inner City
Extreme Affluent Suburb
Extreme Rural
Other (all other STOC categories--see chapter 3)

Color

Black
White

Sex

Male
Female

Parental Education

No High School
Some High School
Post High School

Region

Northeast
Southeast
Central
West

National

Statistical Procedures

P-values. We compute p-values for each of the groups listed above. A p-value is the proportion of respondents within a group who gave an acceptable response to an exercise; that is, it is the number of respondents in a group who gave an acceptable response, divided by the total number of

---

[2]These groups are defined in chapter 6.

respondents in the group.[3] Suppose, for example, there are 1500 respondents in a group; and, of these, 650 gave an acceptable response. The p-value for the goup would be:

# of acceptable responses	Total # of responses	=	p-value
650	1500		.433

Arcsine - transformation. P-values are easily interpreted, but they possess unwieldy mathematical characteristics when they are used to construct an index. Therefore, the p-values were converted by the arcsine transformation:

$y = 20 \arcsin \sqrt{p} - 15.71$ where:

p = the p-value for some group; and

y = the transformed p-value measured in radians with a possible range of -15.71 to +15.71.

The arcsine transformation for the p-value computed in the example above would be:

$y = 20 \arcsin \sqrt{p} - 15.71$

$y = 20 \arcsin \sqrt{.433} - 15.71$

$y = 20 \arcsin (.658) - 15.71$

$y = -1.33$

The value of y is positive when the p-value is larger than .50, negative when the p-value is smaller than .50, and exactly zero for a p-value of .50.

Effects. Once the original p-value have been transformed into y-values, we form effects by subtracting the National y value from the y-value effects for each of the groups used to construct the selection index (see p. 68).

---

[3]A p-value multiplied by 100 gives the equivalent percentage of success (see chapter 8).

Exhibit 9-1

Effects Used in the Selection Index

Index Label	Effects
I	y(Extreme Inner City) - y(National)
S	y(Extreme Affluent Suburb) - y(National)
R	y(Extreme Rural) - y (National)
O	(Other TOC) - y(National)
M	y(Male) - y(National)
F	y(Female) - y(National)
W	y(White) - y(National)
B	y(Black) - y(National)
NHS	y(No High School) - y(National)
SHS	y(Some High School) - y(National)
PHS	y(Post High School) - y(National)
NE	y(Northeast) - y(National)
SE	y(Southeast) - y(National)
C	y(Central) - y(National)
W	y(West) - y(National)

Suppose that on the same exercise for which we computed a hypothetical p-value and arcsine transformation, the National p-value is .39. The corresponding y-value is -2.15. The effects, therefore would be:

y(group)	-	y(National)	=	Effect
-1.33	(-)	-2.15	=	+0.82

The value of an effect is positive if the group p-value is above the National p-value and negative if the group p-value is below the National p-value.

The selection index. The selection index is a linear combination of effects for each exercise. We choose the effects to be used in the selection index (SI) on the basis of their variabilities and their correlations. If an effect has a limited range of values (Low variability), all the p-values for a group are nearly identical for all exercises; therefore, we need not be concerned about the released and unreleased exercises being different with respect to that group. Similarly, if two effects are highly correlated, it may suffice to consider only one of them since they provide redundant information. Based upon these criteria, the selection indices

for Reading and Literature differed somewhat. For Reading:

$$SI = -I + S + W - B - NHS + PHS + NE - SE.$$[4]

For Literature:

$$SI = - I + S - B - NHS + PHS - SE.$$

Selection of Reading Exercises

The computer program for the random selection of exercises based upon the selection index was designed to provide that the entire range of p-values is represented for the nation as a whole and for all the groups. It can happen, however, that some group p-values are biased toward the high or low end of the scale. In reporting, this could give not only a distorted current picture of a group's ability, but--when the subject area is reassessed--it could give a false impression of progress or regression. Although we are currently reporting one-half of the Reading exercises, we selected only one-third of the exercises to be reported by the random procedure and left the remaining one-sixth to be selected systematically. In this way, the reported exercises represent the entire range of p-values for all groups.

The Reading exercises were initially grouped according to objective. The exercises within each objective were ordered by their selection indices from those having the largest positive values through zero to those having the largest negative values; for example, +2, +1, 0, -1, -2. The computer then randomly selected one-third of the exercises. This was done by randomly designating one exercise from every cluster of three exercises throughout each objective-distribution of exercises. Suppose, for example, there are 12 exercises in an objective and we want to select one-third of them. The distribution of the exercises by selection index would be as shown in Exhibit 9-2.

Note that the selection indices descent in value and that one exercise in each cluster of three is starred indicating that it has been selected for release by the random procedure programmed on the computer.

---

[4]These contrasts are defined in Exhibit 9-1.

Exhibit 9-2

Distribution of Selection Indices and One-Third Selection of Exercises

Cluster	Exercise #	Selection Index	Selected
I	NO1-05	11.36	
	NO3-11	10.97	*
	N11-01	9.86	
II	NO-02	6.33	
	NO1-07	6.32	
	NO8-12	5.98	*
III	NO4-11	4.41	*
	N11-02	3.92	
	NO2-11	1.76	
IV	NO2-05	0.72	*
	NO11-06	0.19	
	NO3-14	-0.16	

Again consider the same 12 exercises we used in the random selection example above. Suppose that when we examined the Northeast contrasts we found that the four randomly selected exercises represented only the relatively large Northeast effects as shown in Exhibit 9-3. Exhibit 9-3 shows that the four randomly selected exercises represent only the upper half of the total range of the Northeast abilities assessed by the 12 exercises. In order to obtain a more accurate representation of the entire spectrum of Northeast abilities, the set of randomly selected exercises must be systematically augmented to include some exercises representing the poorer Northeast abilities. For example, by selecting exercises NO1-07 and NO2-13 or NO3-14, we have our desired 50% of the exercises selected and the entire Northeast range of abilities is represented.

To the extent that exercises at different age levels were unique, the random-systematic selection procedure was executed separately for the age levels. Some exercises were administered

at more than one age level. If such an exercise were selected for release at one age level, it would automatically be released at any other age level to which it was administerd. We are reporting all of a limited number of exercises concerned with skimming and scanning or reading rate.

Exhibit 9-3

Distribution of Northeast Contrasts and One-Third Selection of Exercises

Exercise #	NE Contrast	Selected
N03-11	2.67	*
N11-01	2.63	
N08-12	2.55	*
N02-05	2.43	*
N01-05	2.36	
N04-11	2.19	*
N02-02	1.96	
N11-02	1.87	
N01-07	1.53	
N11-06	1.32	
N02-13	1.09	
N03-14	0.76	

Selection of Literature Exercises

The Literature exercises were grouped first by objective, then by exercise format within objective. Finally, within this structure, exercises were grouped by the age level or the combination of age levels (when administered at more than one age level) at which they were administered. The exercises within this hierarchical grouping were ordered by their average (mean) selection index for all ages from those having the largest positive values through zero to those having the largest negative values; for example +2, +1, 0,-1, -2. The computer then randomly selected one-half of the exercises in the pool by designating one exercise from every cluster of two exercises in each hierarchical grouping. Suppose, for example, there are 12 exercises in such a grouping and we want to select one-half of them. The distribution of the exercises by average selection index ($\overline{SI}$) would be shown in Exhibit 9-4. Note that the average selection index descends in value and that one exercise in each cluster of two has been starred indicating that it has been selected for release by the random procedure programmed on on the computer. No systematic selection was done for Literature. All exercises which involved direct interview and/or tape recorded responses are being reported.

Exhibit 9-4

Distribution of Average Selection Indices and One-half Selection of Exercises

Cluster	Exercise #	$\overline{SI}$	Selected
I	N01-05	11.36	*
	N03-11	10.97	
II	N11-01	9.68	
	N02-02	6.33	*
III	N01-07	6.32	
	N08-12	5.98	*
IV	N04-11	4.41	*
	N11-02	3.92	
V	N02-11	1.76	*
	N02-05	0.72	
VI	N11-06	0.19	
	N03-14	-0.16	*

CHAPTER 11

OVERVIEW OF INFERRING POPULATION FACTS FROM SAMPLE DATA

Populations and Samples

A population includes all the individuals in some defined group. Examples of some National Assessment populations are: all 9-year-olds living in the Central region of the nation or all 17-year-olds living in the Extreme Inner City. In our reports, we wish to talk about the knowledges, understandings, skills, and attitudes of entire populations. It is usually not practicable, however, to obtain data from entire populations and for National Assessment to do so would be impossible. Therefore, we obtain data from a portion of the total population called a sample. If we select a sample with care in accordance with certain rules, we can say that the sample is representative of the population from which it was selected and, likewise, that the data obtained from the sample are representative of the data we would obtain from the entire population. National Assessment samples have been selected in such a way that they are representative of the populations from which they were selected.[1]

From Sample Data to Population Facts

The advantage gained by obtaining data from samples rather than from populations is somewhat qualified by a loss of precision in the descriptions we can give of populations on the basis of that data. Within the limits of error due to measurement, the data we obtain from a sample precisely describes that sample. Even when we have a sample which is representative of the population from which it was selected, we cannot state with absolute certainty that the data we obtain from it is exactly true for the respective population. This is because there are many potential samples which could have been selected from the same population--all selected with equal care and in accordance with such rules that each one would be representative of that

---

[1]Chapter 4 gives a detailed description of the National Assessment sampling procedures.

population. Even with all this care, one would not expect to obtain exactly the same percentages of success or group effects from all these samples. The variation that would occur among the percentages of success or among the group effects obtained from all the potential samples which could have been selected from the same population is called sampling error.

We can compute a statistic called the standard error for any percentage of success or group effect. A standard error is an estimate of the variation that would occur among the percentages of success or among the group effects for all potential samples that could have been selected from the same population--the larger the standard error the greater the variation. The larger the standard error for a percentage of success or a group effect, the less precise is the statement we can make about the population percentage of success or group effect.

* * * * * * * *

The details of the development of the limits discussed above and their rationale are given in chapter 12.

CHAPTER 12

DETAILS OF INFERRING POPULATION FACTS FROM SAMPLE DATA

In chapter 11, we defined populations and samples and explained why it is necessary for National Assessment to obtain data from samples rather than entire populations. We also noted that there is a loss of precision when we describe population facts on the basis of data obtained from samples. This chapter presents the details of and the rationales for the statistical procedures which allow us to infer population facts from sample data and to describe the degree of imprecision involved.

Sampling Error

Whenever we infer population facts from sample data, we must bear in mind that a very large number of samples could be selected from the same population--all selected with equal care and in accordance with such rules that each one would be representative of the population from which it was selected. Even with all this care, we would not expect to obtain exactly the same percentages of success or exactly the same group effects from all these potential samples. The variation that would occur among the percentages of success or among the group effects for all these potential samples is called sampling error. The concept of sampling error is important when we infer population percentages of success and group effects from sample percentages of success and group effects. Because of sampling error, we cannot state with absolute certainty that the value of the population percentage of success or group effect is exactly the same as the value of the obtained sample percentage of success or group effect. Most of the percentages of success and group effects for all potential samples would be quite close to the population percent of success and group effect, but a few would differ by a large degree.

In order to see more clearly just how this works, let us consider a bag containing 25 marbles--20 blue marbles and five red marbles. The 25 marbles are the population and the population facts are: there are 80% (or 20) blue marbles and 20% (or five) red marbles. Suppose, however, that we do not know the population facts and want to estimate them on the basis of data obtained from a sample of five marbles selected from the

population of 25 marbles. There are 53,130 possible samples of five marbles that could be selected. If we select a sample of five marbles in such a way that each of the 53,130 possible samples has an equal probability of being selected, we can state that our sample is representative of the population. Some of the possible samples would have 80% (or four) blue marbles and 20% (or one) red marble--the exact same percentages as the population. Other samples, however, would have various other percentages of blue and red marbles. Therefore, we cannot state with absolute certainty that the percentages of blue and red marbles in the population is exactly the same as the percentages of blue and red marbles we obtain from a single selected sample. The differences between the percentages of blue and red marbles obtained from the sample and the true percentages of blue and red marbles in the population is sampling error.

This example with marbles shows how sampling error can occur, even when a sample is representative of the population from which it was selected. In the marble example, we knew the population facts; but in the case of National Assessment percentages of success and group effects, we do not know the population facts. Concommitantly, we do not know the extent or sampling error for any given sample percentage of success of group effect. It is for this reason, we cannot make exact statements about population facts on the basis of sample data.

We can compute a statistic called the standard error for any sample percentage of success or group effect. A standard error is an estimate of the variation that would occur among the percentages of success or among the group effects for all potential samples that could be selected from the same population. We use the standard error and other statistical conventions to make statements about population percentages of success and group effects in terms of probabilities. The details of the procedures for inferring population percentages of success for sample percentages of success and for inferring population group effects from sample group effects are given in the following sections.

For Sample Percentages of Success to Population Percentages of Success

When we infer a population percentage of success from a sample percentage of success, we use the standard error to establish a range or an interval of percentages of success--called a confidence interval--which we are reasonably certain

will include the population percentage of success. The problem lies in the definition of reasonably; i.e., how confident should we be? The answer to this question resides in two risks accompanying the establishment of a confidence interval which probably includes a population percentage of success. If we establish a confidence interval based upon a degree of confidence approaching certainty, the size of the confidence interval could be so large as to be useless. On the other hand, if we establish a confidence interval based upon a low degree of confidence, the size of the confidence interval would tend to be small; but there would be a large probability that the interval is not wide enough to include the population percentage of success. We need to compromise these two risks so that, while we can be reasonably confident that the interval includes the population percentage of success occurs, the interval is not so large that it tells us little about the value of the population percentage of success. The size of any given confidence interval depends upon the value of the standard error and the degree of confidence with which we wish to state that our interval includes the population percentage of success.

This is best shown by considering three examples using hypothetical data. In all three examples, we use exactly the same sample percentage of success to illustrate that inferring a population percentage of success from a sample percentage of success is not a simple process.

Example 12-1. Let us assume we have the following sample data:

Percentage of Success:	70%
Standard Error:	1%

We wish to establish a confidence interval which we can state with 95% confidence includes the population percentage of success. We use the statistical convention that when we have a large number of observations, 95% of those observations occur within 1.96 standard error units on either side of the mean (average).[1] Our confidence interval would be between 70% -(1.96 x 1%) and 70% + (1.96 x 1%) or between 68% and 72%.[2]

---

[1] A sample percentage of success is an estimate of the mean percentage of success for all potential samples that could have been selected from the same population.

[2] Final percents are rounded to the nearest whole percent.

Therefore, we can state with 95% confidence that the interval 68%--72% includes the population percentage of success.[3]

Example 12-2. Let us assume we have the following sample data:

Percentage of Success:	70%
Standard Error:	3%

Again, we wish to establish a confidence interval which we can be 95% certain includes the population percentage of success. We use the same statistical convention to establish the confidence interval as we used in example 12-1. Our confidence interval would be between 70% -(1.96 x 3%) and 70% +(1.96 x 3%) or between 64% and 76%. Therefore, we can state with 95% confidence that the interval 64%--76% includes the population percentage of success.

Example 12-3. Let us assume we have the following sample data:

Percentage of Success:	70%
Standard Error:	1%

These are the same data we used for Example 12-1; but for Example 12-3, we wish to establish a confidence interval which we can state with 99.8% confidence includes the population percentage of success. We use the statistical convention that when we have a large number of observations, 99.8% of those observations occur within 3.09 standard error units on either side of the mean. Our confidence interval would be between 70% -(3.09 x 1%) and 70% +(3.09 x 1%) or between 67% and 73%. Therefore, we can state with 99.8% confidence that the interval 67%--73% includes the population percentage of success.

Summary. Because of sampling error, we cannot infer that a population percentage of success is exactly the same as the

---

[3] If we select 100 samples and establish confidence intervals around the 100 obtained percentages of success in the manner described, on the average, 95 of them would include the population percentage of success.

percentage of success obtained from the sample. We can, however, use the standard error and statistical conventions to establish a confidence interval which we can be reasonably certain includes the population percentage of success. On the basis of the three examples, we can generate two rules regarding the size of confidence intervals.

1. When a sample percentage of success has a large standard error, the size of the confidence interval must be larger than when the sample percentage of success has a small standard error in order to state with an equal degree of confidence that the interval includes the population percentage of success. Comparing Example 12-1 with Example 12-2, note that in order to state with 95% confidence that the population percentage of success occurs within the interval we need an interval three times as large for Example 12-2 as for Example 12-1.

2. For any given standard error of a percentage of success, the size of the confidence interval must be larger when we wish to state with a high degree of confidence than when we wish to state with a lower degree of confidence that the population percentage of success occurs within the interval. In comparing Example 12-1 with Example 12-3, note that both examples have the same standard error, but the 99.8% confidence interval (Example 12-3) is one and one half times as large as the 95% confidence interval.

A confidence interval should not be regarded as a mystical contrivance used solely by statisticians; we encounter them in everyday life--although not in the strictly formal sense we have used here--when it is not possible to make precise statements. Suppose, for example, that you order an article of merchandise and the salesman tells you that the article will be delivered in 10 days give or take two days. The salesman has established a confidence interval--the give or take two days--around tne average time it has taken that article to be delivered in the past and within which he is reasonably confident the article will be delivered this time. The confidence intervals we establish for percentages of success are much the same.

From Sample Effects to Populati n Effects

If a group population were compared to the population for the nation as a whole, we would never expect to obtain an effect exactly equal to zero; i.e., some difference, however small, would exist. Many population effects, however, would be so small that they can be called equal to zero for practical consideration.

When we infer a population effect from a sample effect, we are concerned, first of all, with three considerations.

1. Is a sample effect large enough that we can be reasonably certain that the population effect is not equal to zero; i.e., does a real difference exist between the group and the nation as a whole?

2. Is the sample effect small enough that we can be reasonably certain that the population effect is, for practical consideration, equal to zero; i.e., does no difference of practical magnitude exist between the group and the nation as a whole?

3. Is the sample effect not large enough that we can be reasonably certain that the population effect is not equal to zero and not small enough that we can be reasonably certain that the population effect is, for practical consideration, equal to zero; i.e., can we make no reliable statement as to whether a real difference exists or whether no practical difference exists between a group and the nation as a whole?

When we infer a population effect from a sample effect, we are concerned, secondly, with the risk of making two types of error which are related to the three considerations stated above. The first type of error is calling a population effect not equal to zero when, in fact, the population effect is, for practical consideration, equal to zero. The second type of error is calling a population effect, for practical consideration, equal to zero when, in fact the population effect is not equal to zero. We can never eliminate the risk of making both types of error--or even one type of error--completely.

CHAPTER 13

THEMES

Beginning with the year 02 assessment (Reading and Literature), we are reporting the results for each subject area in a series of theme reports.[1] Each report presents the results a group of exercises which share a common theme or unifying idea. For example, one of the Literature themes is "Understanding the language of literature." A report volume presents the results for a variety of different exercises which were designed to assess an individual's knowledge of the way language operates in poetry and prose. In Reading, one of the themes is "Read for main ideas and organization of passages." This theme embodies exercises which require individuals to read a passage and detect its central thought or topic or to detect the mode in which the material is organized.

The concept of themes arose from the desire of National Assessment to report its results in a manner meaningful to educators and other concerned persons. If we divide the entire pool of exercises for a subject area into sets of exercises which have a common theme, educators should be able to answer such questions as the following:

1. Do various groups perform differentially in a given situation?

2. How does a given group perform in one situation as opposed to other situations?

3. Does a group perform better on certain aspects of a given situation than on other aspects of the same situation, and are these aspects identifiable.[2]

---

[1]See Foreword, the section--Reporting Format.

[2]The themes for Reading and Literature are outlined and discussed briefly in the Reading Summary and Literature Summary, respectively. Each theme is discussed in detail in the respective theme volumes.

The themes for both Reading and Literature were developed post hoc; that is, they are based on only those exercises which were actually administered. These themes do not exhaust all possible contents in the respective subject areas since the exercises which were developed for these subject areas were not, themselves, exhaustive.

In each subject area, the pool of administered exercises was examined critically by several National Assessment staff members who categorized the exercises into themes. The pool of exercises was simultaneously sent to a consultant (a scholar in each respective subject area), who also examined the exercises and categorized them into themes. When both the staff members and the consultant had arrived--independently--at reasonable sets of themes, the staff members and the consultant met to resolve differences and to finalize a "consensus" set of themes.

Once viable themes have been developed for a subject area, they serve two somewhat related functions. First, each of them serves as the content of a report volume. Second and more important, theme reports allow educators and other concerned persons to examine in a single volume the results of a meaningful set of exercises. Each theme volume gives results for the nation as a whole and the various groups defined in chapter 3. The reader of these reports can examine behavicrs of groups for the theme in general (as expressed by the median[3]) or for specific exercises.

---

[3]The median is the mid-point of the range of percentages of success for all the exercises within a theme. A median can be computed for the nation as a whole and for each group.

CHAPTER 14

THE MERITS AND WEAKNESSES OF ADJUSTMENT (INCLUDING BALANCING)

The educational administrator wants to make comparisons between groups, to find out who is learning more and who less in hopes of being able to improve performance in the lagging groups. Indeed, he would like to go further and find out what factors to change and how much changes in these factors would strengthen the educational achievement of the students affected. For example, when we find that boys know less about the reproductive system of both sexes than do girls, this raises at once the question of strengthening the education of boys. Inevitably the desire is to subdivide the country into finer and finer groups so as to make comparisons between subgroups that have "everything alike" except the variable being studied.

In other words, we search for causes of the differences. Unfortunately, we cannot have "everything alike" in social problems and rarely in physical problems either, and so we are not actually able to carry out the precise programs. But half a loaf may be better than none, and so we may carry out that part of the program that seems feasible. We subdivide by important variables and make comparisons in performance among groups.

One thing that happens is that as we introduce several variables the number of subdivisions grows like a product. For example, if we have 5 variables with 2, 3, 5, 7, and 4 categories respectively, we have $2 \times 3 \times 5 \times 7 \times 4 = 840$ subgroups, and a sample of 8400 people would give an average of only 10 per subgroup. Naturally many subgroups would be empty and many fuller than 10, but it will still be hard, if not impossible, to make comparisons among subgroups, for some will be too sparse.

We might try to avoid these sparse cells by only looking at factors one at a time.

However, children in the Extreme Affluent Suburb tend, more than children in the Extreme Inner City, to have better educated parents. Because of this lack of balance, part of the difference between these two groups may be considered as growing out of the difference in parental education.

It is natural to ask, "What would the difference between these extreme types of community have been if the distribution of parental education, sex, color and region had been the same

for both types of community referred to above?" Were it possible to rearrange the world to equate these distributions for each type of community, the effects upon our nation and its schools would be profound. Such rearrangement is not possible. It is usually appropriate to think of the balanced results presented in a later special report as reflecting the differences we would see in the absence of masquerading by the other four factors. We can be reasonably sure the balanced results do a much better job than the unadjusted results of reflecting such differences.

Still another question concerns the combination of factors. The performance of a given group may be found to differ, depending upon subgroupings on other variables. Thus, the effect associated with Extreme Affluent Suburbs may be different in the Northeast and the Southeast. Or the effect associated with sex may be somewhat different for Blacks and Whites. Such interactive differences may be of importance; balancing does not adjust for them.

It is natural to ask whether this or any such method of analysis can help us. To some extent they can aid, to some extent not. We cannot make up for cases we don't have but we may be able to supply approximate analyses that will come near to answering such a question as what is the effect of region of the country on performance when you control for size-and-type of community and several other variables. If the effect of region is substantially reduced by the analytical adjustment, we may be inclined to think that region is not in itself the cause of the raw differences as much as the other variables. One role of adjustment then is to help us make approximate comparisons and summaries that we cannot make by directly subdividing all the variables.

Elsewhere (see Foreword and chapter 8) we have many cautionary remarks about the dangers of misinterpreting the causative powers of given background variables, for they may be poorly measured and they may not mean what they say. For an example from the field of warfare, in World War II the more fighter opposition bombers had, the closer to the target were the bombs. Why? Fighters didn't come up when the weather obscured the target. Such proxy variables, especially when their correct interpretation may be the absolute reverse of their obvious effect puts us in grave danger of making mistakes. We do not go further into that here.

Nothing but experimentation, if that, can serve to demonstrate what the actual effect of changes will be. We are, however, trying to get hints and insights from the data we have. Furthermore, if someone does have a causative model involving the variables National Assessment measures, he does have a chance to check it against these results.

We see then that the purpose of analysis and adjustment is to help the data reveal information that they cannot give in their raw form. Aside from the dangers of misinterpretation, we have the political arguments for and against adjustment. First, against: if adjustment for background variables seems to reduce the differences between a group of the population and the national average, it has been argued that this tends to minimize the disadvantage of the group and, it is further argued, that adjustment should not be made. The direction of the effect of an adjustment is not necessarily one-way; adjustments can increase differences as well as decrease them. Those arguing against adjustment in the reduction case would presumably argue for it in the case of increased discrepancies.

A second argument favors adjustment. It argues that we must adjust for important variables (presuming that the adjustment will reduce effects) so that we show the potential of the disadvantaged group.

Clearly the people making the first and second argument want the same thing, to improve the position of the disadvantaged group, and of course, this is a national goal. Steps toward achieving such goals do depend on searching for causes and methods of improvement, on finding weak spots in a system and so on. We should, therefore, look at our data in every way we can for hints about how the system works and how to improve it. Analysis and adjustment are tools for doing this. The question is not whether to adjust or not, but, "What are the useful ways?" and, "What do the variables mean?", "What further variables do we need to measure?", and "How shall we interpret the results?".